MODELS FOR HIERARCHICAL-STRUCTURED ITEM RESPONSE DATA AND A LONGITUDINAL MULTILEVEL LOGISTIC REGRESSION MODEL ON DIF ANALYSES

A Dissertation

by

XUEYING HU FRANCIS

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Bruce Thompson
Co-Chair of Committee,	Victor Willson
Committee Members,	Prathiba Natesan
	Ping Xiang
Head of Department,	Victor Willson

August 2015

Major Subject: Educational Psychology

ABSTRACT

The presented journal article formatted dissertation investigated the performance of two models for hierarchical-structured item response data (i.e., Kamata's MLIRT model, and Multiple Regression IRT model) and discussed an application of the multilevel IRT modeling, i.e., a longitudinal multilevel logistic regression model for DIF analyses. Study I compared the estimates of abilities and IRT difficulty parameters of the two models for multilevel-structured IRT data. *Bias* and *RMSE* were compared under 8 conditions (2 test lengths, 4 intraclass correlation coefficients, i.e., *ICC*). Study II sought to learn the causes of DIF, specifically investigating if DIF arises through higher-level clusters, such as different schools, and longitudinal sources, such as multiple time points of test (e.g., beginning *v.s.* end of year). The accuracies of DIF detection at each level were evaluated under 48 conditions (2 test lengths, 2 percentages of DIF items at school-level, 2 percentages of DIF items of time-level, 3 sample sizes, 2 magnitude of DIF) by power and Type I error rate.

Findings of Study I provided guidelines for model selection between MLIRT and MR-IRT model. Results indicated more accurate estimates of school abilities but less accurate estimates of student abilities with MLIRT model. MR-IRT was found more appropriate to use when sample size was small. For both the MLIRT and MR-IRT models, the longer test length resulted in more accurate estimates. *ICC* played an important role in estimating the school variances of abilities. Study II examined the power and Type I error of DIF detection with the proposed model. Results showed an

overall powerful DIF detection. Type I error rates at each level roughly fell into the liberal range of Bradley (1978), 0.025 to 0.075. Consistent with previous study, the magnitude of DIF at each level and the sample was found to be the most important factors for a powerful DIF detection. In general, the time-level detection had higher power than the school-level.

DEDICATION

This dissertation is dedicated to my family for their love and supports throughout all the years. First and foremost to my parents, Hua Li and Xiaosong Hu, for their always encouragement and ever-lasting love. They raise me to a grateful, honest, and independent woman that I am today; they provided and continue to provide me the best they can; they always encourage me to chase my dreams, no matter how unrealistic they seem to be. I will never reach this far without them.

To my mother-in-law, Ginny Francis and my father-in-law, Harry Francis, for their supports through words and actions, their love and their confidence in me. My mother-in-law, Ginny Francis, set me the best example of a loving, faithful, passionate, creative and strong woman. Her words "you can do it" will always be with me and keep me to move forward.

To my sister-in-law, Alison Francis, for her encouragement, understanding, and constant love and support.

Finally, to my most loving and supportive husband, Michael Francis, I thank you for always being here for me, in good days and bad days. Thank you for having faith in me even when I started to lose it in myself. Thank you for your hugs, your caring, your humors to cheer me up, and your understanding when I am frustrated. Things are simply better when I share them with you. You are the equal part of this achievement. I love you with all my heart and I am so grateful to have you in my life.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Bruce Thompson and my co-chair, Dr. Victor Willson, for their support, guidance, understanding, and patience from the beginning to the end of my PhD study. I am extremely grateful to have Dr. Thompson and Dr. Willson, who allow me to explore my own research interests and also provide me with all kinds of supports. My deep gratitude also goes to Dr. Prathiba Natesan, who patiently listened to my research ideas, gave me advice and helped me sort out the technical details of my work. I am also gratefully acknowledging Dr. Ping Xiang, who has always been there to listen and give me most valuable advice.

My thanks also go to my friends and schoolmates and the department faculty and staff for making my time at Texas A&M University a great experience. I also want to extend my gratitude to American College Testing Inc. (ACT) and National Center for Assessment, which provided my internship opportunities, and supported me with practical assessment project experience, that inspired me to come up with the research ideas in the dissertation study.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

INTRODUCTION

Research on the identification of effective schools and criteria for measuring effectiveness is a major topic in education. Multilevel analysis takes the class and school variance into consideration and extends the measurement from item characteristics and individual abilities to group-level measurements (Bryk & Raudenbush, 2002; Fox, 2005; Goldstein, 1987; Longford, 1993). Multilevel modeling can be combined with Item Response Theory (IRT) to estimate the effects of multilevel covariates on a latent trait. This amalgamation of the two models allows us to investigate and analyze the covariates that affect person ability instead of simply estimating the latent traits (Maier, 2001). Additionally, this combination provides more accurate estimation of the standard errors of the parameters (Adams, Wilson, & Wu, 1997; Fox, 2005; Maier, 2001, 2002). Multilevel IRT treats ability parameter of an IRT model as the dependent variable of a multilevel regression model (Fox, 2005).

One of the most critical applications of IRT in educational testing field is for detecting items that function differentially. Differential item functioning (DIF), which is considered as a threat to the validity of the test, has been a serious concern in evaluation and assessment (Thissen, Steinberg, & Wainer, 1988, 1993). There have been many DIF detection procedures available with the use of IRT and significance statistical testing. However, the classic DIF detection procedures can only detect the existence of DIF, but cannot explain the causes of DIF. In the case when there are more than one sources of DIF, the classic DIF procedures fail to examine the fundamental reasons of the DIF

occurrence. Therefore, a model that is able to (a) identify the existence of DIF, and (b) explain the sources of DIF is needed to further facilitate the DIF investigation.

The multilevel IRT model, expressed in the multilevel logistic regression model (Adams, Wilson, & Wu, 1997; Kamata, 2001; Mellenbergh, 1994), treats the coefficients that are associated with DIF as random effects and is able to estimate the DIF at each level through the estimation of variances of the random effects (Swanminathan & Rogers, 1990). Multiple causes of DIF could be simultaneously analyzed in one multilevel model.

Study Number One

This dissertation consisted of two studies. The first study compared the parameter and variance estimate accuracies across two hierarchical item response models. First, 1PL Three-Level MLIRT model ("MLIRT") proposed by Kamata (1998, 2001) was designed to accommodate item responses data that were collected in hierarchical settings and allowed investigation of item response data that contained hierarchical structure (Fox & Glas, 2001; Kamata, 2001; Maier, 2001, 2002). MLIRT combined multilevel modeling that were formulated via hierarchical linear model (HLM) (Bryk & Raudenbush, 2002) with Rasch-IRT model to estimate the latent traits, item parameters, and the variations of individuals' performance across groups. Pastor (2003) suggested four advantages of MLIRT model: It allows (a) to treat a latent trait as the dependent variable in a hierarchical-structured data analysis, (b) the dependency found in the hierarchical data, (c) estimation of latent traits at different levels (e.g. student-level, and school-level), and (d) more accurate estimation of the relationships

between independent variables and latent traits (i.e., dependent variables) across levels. However, because MLIRT integrates HLM and IRT, the assumptions of both models need to be met.

The second model uses a two-step Multiple Regression (MR) approach to investigate hierarchically structured data (Gelman & Hill, 2007). The two-step MR approach was combined with the 1PL-Rasch IRT to estimate latent traits in multiple levels and therefore was named the "MR-IRT" model. MR-IRT model was conducted in two steps: (a) estimating IRT parameters with 1PL-Rasch IRT model and (b) using the latent trait estimates, i.e., person ability estimates, from step one, as dependent variable in a MR analysis (Pastor, 2003). MR-IRT is easy to implement as both IRT and MR procedures can be realized by many statistical software packages. The other advantage of this model is that no group-level assumptions need to be met when estimating the IRT parameters, because the IRT is conducted in the single level model. However, the assumption of independence of the observation is almost always violated in hierarchical-structured data (Hox, 2010), which leads to inaccurate estimate of group-level ability variance.

A hierarchically structured item response data was simulated and estimated with both Kamata's MLIRT model and MR-IRT model. The model estimation accuracy was evaluated for various test lengths, sample sizes and intraclass correlation conditions through a simulation study. The simulated hierarchically structured item response pool was generated with SAS 9.3 and the parameter estimates were obtained using Mplus 7.1.

Bias, the root mean square error (*RMSE*), and the standard error (*SE*) of variance were calculated to assess estimation accuracy.

Study Number Two

The second study was an application of MLIRT to detect and explain DIF when measuring a certain trait across groups (Kamata & Binici, 2003) and time points. A multilevel longitudinal logistic regression model was designed for DIF analysis. The longitudinal multilevel DIF detection procedure was aimed to identify (a) the cross-sectional sources of DIF in the student-, class- and school- levels and (b) the longitudinal sources of DIF in multiple time points. In addition, the second paper studied the effects of proportion of DIF items, test length, sample size on DIF detection with the proposed model. Last, the effect of the model misspecification was examined.

The longitudinal multilevel logistic regression model was design based on Kamata's MLIRT model. The multilevel logistic regression DIF model was a hybrid of the logistic regression DIF method (Swaminathan & Rogers, 1990) and multilevel logistic regression IRT model (Kamata, 2001). In the simulation design of study II, only a Rasch IRT model was assumed for DIF analysis (i.e., item discrimination parameter was assumed to be equal across groups). The proposed model was specified and estimated with SAS PROC GLIMMIX (Pan, 2008; Zhu, Rupp & Gao, 2011). The accuracies of item difficulty parameter were evaluated by the *RMSE* and *bias*. The accuracies of DIF detection at each level were evaluated under a variety of conditions by power and Type I error rates (Kristjansson et al., 2005; Pan, 2008).

Organization of Document

This document is divided into four distinct sections. While the first section is the general overview of the document, the latter two sections are organized as two individual journal articles on two related research topics. The last section served as a summary of the conclusions of the two studies. A description of each section is provided as follows:

i. The first section is an introductory section which provides the overview of the research topics on this dissertation.

ii. The second section is an individual manuscript regarding a comparison of two multilevel IRT (MLIRT) models in terms of the estimation accuracy of item and person parameters. This section presents the literature review, methods, results, discussion and conclusion sections of the first journal article.

iii. The third section is another individual manuscript regarding longitudinal multilevel differential item functioning (DIF) estimated with longitudinal multilevel logistic regression model. This section presents the literature review, methods, results, discussion and conclusion section of the second journal article.

iv. The fourth section is a conclusion section which stated the connections of the two studies, and summarizes the significance and findings of each study.

A COMPARISON OF TWO MODELS FOR HIERARCHICAL-STRUCTURED ITEM RESPONSE DATA

In the recent years, evaluating the students' abilities and holding schools accountable for students' performance through high-stakes testing is of great interest in educational assessment field. Unfortunately, the traditional single-level measurement can only estimate the students' abilities, but not able to serve the goal of comparing accountabilities within and between schools. The present study focuses on accurately evaluating the examinees' abilities when they are from different schools with the multilevel IRT modeling. The multilevel IRT model estimates the relationships among the dependent variables at different levels, such as student test scores that are nested within multiple school characteristics. Ignoring the hierarchical structure in our analyses can cause some problems (Burstein, 1980; Cronbach, 1976), such as:

(a) Failure to see some import phenomena such as cross-level interaction effects.

(b) Failure to separate variance components of different levels (Cohen, Cohen, West, & Aiken, 2003, chap. 4), which results in a violation of the homoscedasticity assumption of statistical inference test. The high intraclass correlation between different levels leads to an underestimation of the standard errors of the fixed parameters, which leads to the spuriously "significant" results (Hox, 2010).

Theoretical Framework

Comprehensive literature review was conducted on (a) comparison of multiple regression (MR) model and HLM model, (b) mathematical expression of MR and HLM

models in a school setting, (c) Item Response Theory (IRT), (d) multilevel IRT modeling, including Kamata's MLIRT model, two-level IRT and three-level IRT, (e) MR-IRT model, and (f) Monte Carlo Simulation study.

Comparison of Multiple Regression (MR) Model and HLM Model

Multilevel models are extensions of multiple regression (MR) (Gelman & Hill, 2008) and particularly accommodates data structured in groups. Hierarchical Linear Modeling (HLM), also known as random coefficient models (Kreft & de Leeuw, 1998), variance component models (Longford, 1989), or multilevel random coefficient models (Nezlek, 2001, 2003), takes the random effects and multilevel variance into account. HLM allows analyses of multiple levels simultaneously with the use of a statistical model that includes various dependencies (Hox, 2010). Gelman and Hill (2008) reviewed HLM model and a two-step multiple regression (MR) model.

(a) The two-step MR was described as the "individual-level regression with cluster indicators, followed by cluster-level regression of the estimated cluster effects" (Gelman & Hill, 2007, p. 240). The first step is an individual-level regression, which fits the outcome variable on the individual predictors and the grouping indicators for all groups that are coded as dummy variables. In order to obtain model identification, the constant term in the multiple regression is not included so that the indicators for all schools can be included (Gelman & Hill, 2007, p. 68). The second-step is a group-level regression to estimate the group effects. In this step, the estimated coefficients of the grouping indicators from the previous step are considered as the outcome variable data,

which are then regressed on the group-level predictors. After the two steps, each of the predictors in the model is then included in one of the two regressions.

(b) The HLM is similar to the two-step MR model except that both steps are performed simultaneously. The individual-level outcome variable is predicated by individual-level predictors. The individual-level intercept and coefficients are assumed to vary by groups. In the group-level, the individual-level intercept and coefficients are regressed on the group-level predictors and are decomposed into the group-level mean and standard deviation of the unexplained group-level errors. In HLM, instead of estimating the within-group coefficients, the mean and variance of the coefficients among the groups are estimated to describe the distribution of the coefficients. (Van der Leeden, 1998). The assumption of varying coefficients is what made the model "multilevel". The "varying coefficient" is called "random effects", indicating the randomness in the probability model for the group-level coefficients.

Multiple regression (MR) and HLM approaches coincide when (a) the group-level variation is small, so that the multilevel model reduces to multiple regression model as there is no group indicator any more, (b) when the variation among the group-level coefficients is large (compared to their standard errors of estimation), so that multilevel modeling reduces to MR with group indicators, or (c) when there is very few groups, so that there is too little information to accurately estimate the group-level variation (Gelman & Hill, 2008, p. 247). In these cases, multilevel modeling gains little beyond the MR varying-coefficient models (i.e., with only fixed effects).

The critical difference between HLM and multiple-step MR with OLS estimates is to model coefficients as random or fixed (Raudenbush & Bryk, 2002). Random coefficients are assumed to vary across the higher-level groups, whereas fixed coefficients are assumed to be either constant or vary systematically across groups. While HLM assumes random coefficients and estimates the error terms, the MR models assume that there is no error variance among groups and no variance of error is estimated (Richter, 2006). Therefore, MR is used when only one level has random effects or the error variances in each level are explained in multiple steps rather than simultaneously.

However, if the objective of the research is to generalize the estimates at each level to larger populations from which the samples are randomly selected, the analyses need to be conducted based upon multiple sources of error variance simultaneously in a given study (Richter, 2006). The MR model, as stated, separates the variance components of each level sequentially, rather than simultaneously. Thus, HLM is more appropriate for multilevel parameter estimation based upon multiple sources of error variance.

Additionally, OLS estimation method could not be applied to estimate the variance of error in the MR model with random effects. HLM, however, involves a more complex error structure and applies the ML techniques, instead of OLS, to estimate the variance of error at each level simultaneously (Richter, 2006).

Furthermore, in a hierarchical-structured dataset, the assumption of independence of observations is almost always violated, which leads to underestimate of the standard

errors and spuriously "significant" results when using MR models (Fox, 2005). On the other hand, HLM allows the dependency of observations at each level and therefore is more appropriate to use when the assumption of independence of the observations is violated in the hierarchically structured dataset.

Due to the differences between these two models, the parameter estimation results were expected to be different from the HLM model and the MR model, especially when the effects had to be modeled as random. The estimates of the MR model were expected to be less accurate due to the restriction of modeling only fixed effects and the limitation of estimating variances sequentially rather than simultaneously (Richter, 2006).

Mathematical Expression of MR and HLM Models in a School Setting

HLM was first introduced in the assessment of school effectiveness by Aitkin and Longford (1989) and started to play an important role in assessing the schools' and states' accountabilities as well as students' abilities. Following is the mathematical expression of each model in a school setting.

MR model. Instead of estimating parameters at different levels simultaneously, MR model is conducted in two sequential steps.

Step 1

$$Y_{ij} = \underbrace{\beta_{1j}X_{1j} + \ldots + \beta_{kj}X_{kj}}_{\text{student-level predictors}} + \underbrace{\beta_1 X_1 + \ldots + \beta_j X_j}_{\text{school-indicators}} + e_{ij},$$

where

$X_{1j} \ldots X_{kj}$ are student-level predictors, $\beta_{1j} \ldots \beta_{kj}$ are effects associated with the student-level predictors. *k* is the number of student-level predictors,

$X_1 \ldots X_j$ are the school indicators, $\beta_1 \ldots \beta_j$ are the estimated coefficients associated with each school. *j* is the number of schools, the estimated Y_{ij}, is the mean score of school *j*,

e_{ij} is assumed to be normally distributed, which is a random effect representing the student-specific residual,

Step 2

$$\beta_j = \gamma_{0j} + \gamma_{1j} S + U_j,$$

where

The estimated coefficients of all school indicators $\beta_1 \ldots \beta_j$ from Step 1 regression are regressed on the school predictors *S*,

γ_{0j} is interpreted as the grand mean across schools,

γ_{10} is the mean effect of the school predictors,

U_j is assumed to be fixed effects representing school-specific residuals in school-level regression.

In Step 2, the estimated coefficients of the group-indicators are used as the data for group-level to explain the across-school variability in these estimate (e.g., Juhaz & Rayner, 2003; Stine-Morrow, Millinder, Pullara, & Herman, 2001; Zwaan, 1994).

HLM. The two-level HLM random intercept model with only student-level predictors is expressed as follows,

Level-1 (student): the observed score of student *i* in school *j* is predicted by the level-1 predictors $X_{1j} \ldots X_{(k-1)j}$ plus a random error e_{ij} that is specific to student *i*.

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{1j} + \ldots + \beta_{(k-1)j} X_{(k-1)j} + e_{ij},$$

where

i represents students within a school; *j* represents schools, *j* = 1, …, *k*,

the intercept β_{0j} represents the predicted mean observed score in school *j* when all the predictors = 0. β_{0j} is assumed to be normally distributed,

the slopes $\beta_{1j},\ldots, \beta_{(k-1)j}$ represent the effects of the level-1 predictors on student observed scores.

Level-2 (school): the mean observed score of school *j* is modeled as the mean of the entire sample γ_{00} (i.e., grand mean), plus a random effect U_{0j} that is specific to school *j*.

$$\beta_{0j} = \gamma_{00} + U_{0j},$$
$$\beta_{1j} = \gamma_{10},$$

…

$$\beta_{(k-1)j} = \gamma_{(k-1)0}.$$

The random intercept β_{0j} is assumed to be normally distributed and is further decomposed in the next highest level into a mean and a variance. The random effects e_{ij} and U_{0j} in the model are assumed to be normally distributed with mean of zero, $e_{ij} \sim N(0, \sigma^2)$ and $U_{0j} \sim N(0, \tau_{00})$, respectively, in which the variance σ^2 is the student-level variance of error whereas τ_{00} is the variance of error at the school level. The random effects and the level-1 predictors are assumed to be uncorrelated.

Item Response Theory (IRT)

Item Response Theory (IRT) has been widely used in educational testing (e.g., Bielinski & Davison, 2001; Edward, 1993; Kulick & Hu, 1989; Le, 1999; Pike, 1990;

Rock, Pollack, & Quinn, 1995; Zwick & Ercikan, 1989). IRT has two main postulates, the abilities and the Item Characteristic Curve (*ICC*) (Cantrell, 1999). The abilities measure the performance of an examinee on a test item. An *ICC* is a frequency polygon or an ogive representing the relationship between the item performance and the examinee's abilities that determine the performance (Cantrell, 1999; Hanbleton & Swaminathan, 1985). An *ICC* reflects the probability of selecting a certain response to an item with respect to the ability of the person (Ostini & Nering, 2006).

The most common IRT model is the one-parameter logistic model, which is also known as the 1PL-Rasch model (Rasch, 1960; Thissen, 1982; Wright, 1977). The item response is binary, with $Y_{ij} = 1$ being a "correct" or positive response and $Y_{ij} = 0$ being an "incorrect" or negative response. This model specifies the probability of a person *j* responding correctly to item *i* ($Y_{ij} = 1$) conditional on the "ability" of the person *j* (θ_i) as

$$P(Y_{ij} = 1 | \theta_i) = \frac{1}{1 + exp[-a(\theta_i - b_j)]}.$$

In the IRT models, abilities (θ_i) are usually assumed to be normally distributed with mean of zero and variance of one. The IRT parameter b_j represents the item difficulty, which determines the location of the logistic curve along the ability scale (Hedeker, Berbaum, & Mermelstein, 2006). The item difficulty is interpreted as the ability of a person that has a 50% probability of answering an item correctly (Chaimongkol, 2005). The normal range of item difficulty parameter is roughly from -3 to +3, with the lower values indicating the easier items and higher values indicating more difficult items. The IRT parameter *a* is item discriminating parameter, which is the

slope of the logistic curve, i.e., *ICC*. The larger the item discriminating parameter, the steeper the *ICC* is. IRT-*a* is the "discriminating" parameter in the sense that it discriminates items by ability θ_i. The regular range of discrimination parameter is 0 to 2. In the Rasch model, all items are assumed to have the same slope, so IRT-*a* is assumed to be constant across the items, and therefore, does not carry the *j* subscript. A more general form of IRT model is the 2PL-IRT, by releasing the restriction of the discrimination parameter in the Rasch IRT model and adding the *j* subscript to IRT-*a* parameter.

The three-parameter logistic (3PL) IRT model is an even more general form of the 2PL model and includes a pseudo-guessing parameter. The 3PL-IRT model is formulated as

$$P_j(\theta) = c_j + (1 - c_j)\frac{exp(a_j(q - b_j))}{1 + exp(a_j(q - b_j))},$$

where c_j is the pseudo-guessing parameter, which is the probability of answering item *i* correctly for persons with very low ability (Chaimongkol, 2005). By adding the pseudo-guessing parameter, the probability of answering item *i* correctly for low ability persons is nonzero, which is a noticeable difference from the 1PL- or 2PL-IRT models. In the present study, only a 1PL-IRT model was used, meaning the only item difficulty parameter *b* was considered in the model (Reise, 2000).

Multilevel IRT Modeling

A multilevel IRT model is a combination of HLM and IRT, where the ability parameter in the item response model is treated as the dependent variable in the higher-

levels of an HLM (Fox, 2005), which allows the estimation of item parameters, students' abilities, and school-characteristics. Not only does multilevel IRT allow the estimates of item parameters and person abilities, multilevel IRT investigates the school-level effects and the interaction of person- and school-level effects as well. The present study aimed to evaluate two multilevel IRT models: the Kamata's Multilevel IRT (MLIRT) model (Kamata, 1998, 2001) and the two-step Multiple-Regression IRT model (MR-IRT).

Kamata's MLIRT Model. The Kamata's MLIRT is essentially an amalgamation of the Rasch IRT model with Hierarchical Linear Model (HLM) that allows investigation of item response data that contain hierarchical structure (Fox & Glas, 2001; Kamata, 2001; Maier, 2001, 2002). The two-level MLIRT framework has the item-level and student-level estimating items parameters and person characteristics, respectively. The two-level framework has been proven to be mathematically equivalent to Rasch IRT model, considering person parameter as the random effect (Kamata, 2001). On top of the two-level framework, the third-level, school-level, is included to take the school variance into account, estimating the school effects and the interaction of student- and school-level effects. As the variances explained in the student-level and the school-level are investigated simultaneously, the MLIRT model ensures a more accurate estimation than the single-level Rasch IRT model. MLIRT involves HLM to estimate the random effects at each level simultaneously, thereby avoiding the need to perform separate analyses (Adams, Wilson, & Wu, 1997; Kamata, 1998).

Although MLIRT and HLM share similar framework, one distinguishable difference between HLM and MLIRT is the distribution of the outcome variables on the

lowest level of the model. Recall that in HLM model, the outcome variable Y_{ij} is assumed to be continuous. Therefore, the multiple linear regression is used for the student-level.

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{1j} + \ldots + \beta_{(k-1)j} X_{(k-1)j} + e_{ij}.$$

Because the outcome variable Y_{ij} is assumed to be normally distributed, the distribution is determined by two parameters, the mean and the variance, as $\beta_{0j} + \beta_{1j} X_{1j} + \ldots + \beta_{(k-1)j} X_{(k-1)j}$ represents the mean of the outcome variable, e_{ij} represents the variance of the dependent variable. In MLIRT model, the lowest level is the item-level, whose outcome variable is dichotomous item response. Unlike multiple regression, there is no error term in this logistic regression model due to the property of the Bernoulli distribution of the outcome variable, $E = P_{ij}$, $Var = P_{ij}(1-P_{ij})$, where P_{ij} is the probability of student *j* endorsing the correct response to item. Once the probability is known, the Bernoulli distribution of the outcome variable is obtained without the need to include the error term. However, the higher levels would have error terms if the outcome variables were continuous variables and they follow the normal distribution. The mathematical expression of the two-level and three-level IRT will now be discussed.

The simplest MLIRT model is the **two-level IRT model** with the items (Level-1) nested within students (Level-2) (Kamata, 1998). The two-level nested model facilitates the estimation of measurement error within and between these two levels. MLIRT models consider the item difficulty (location) as the Level-1 variable and the person ability, attitude, or latent trait as the Level-2 variable (Natesan, 2007).

In the item-level, the logit link function (i.e., the logarithm of odds) is used to linearize the nonlinear relationship between the predictors X_{ij} and the probability P_{ij} and to transform the model to a logistic regression model to meet the boundaries of the probability [0, 1].

$$\eta_{ij} = \log(odds) = \log(\frac{P_{ij}}{1-P_{ij}}),$$

where P_{ij} is the probability of person *j* providing the correct answer to item *i* and η_{ij} is the item response.

A Bernoulli sampling model is used for a dichotomous outcome variable in the item-level, which assumes the independency of all the trials. In other words, all the items are assumed to be independent to each other. The probability of success is P_{ij} while the probability of failure is $1\text{-}P_{ij}$ for item *i* and person *j*.

$$\eta_{ij} = \beta_{0j} + \beta_{1j}X_{1j} + \beta_{2j}X_{2j} + \cdots + \beta_{kj}X_{kj}$$

$$= \beta_{0j} + \sum \beta_{qj}X_{qj},$$

where β_{0j} is the intercept, calculated as the predicted *log(odds)* when all X_{ij}= 0. β_{1j} is the effect associated with item 1, β_{2j} is the effect of item 2, and so on (Natesan, 2007). X_{qj} is the q^{th} dummy variable for person *j*. When $q=i$, X_{qj} is 1, otherwise, X_{qj} is 0. The mean and variance of item responses η_{ij} are P_{ij} and $P_{ij}(1\text{-}P_{ij})$, respectively (Kamata, 2001).

The effect of the k^{th} item, which is the "reference item", is assumed to be zero. Then equation 1 is altered to,

$$\eta_{ij} = \beta_{0j} + \beta_{1j}X_{1j} + \beta_{2j}X_{2j} + \cdots + \beta_{(k-1)j}X_{(k-1)j}.$$

The effect of item *i* is associated with the q^{th} dummy variable. Therefore, equation 4 reduces to

$$\eta_{ij} = \beta_{0j} + \beta_{qj},$$

where β_0 is the intercept and β_{qj} is the specific effect associated to the q^{th} dummy variable,

Combining equation 1 and 5, we get

$$\log(odds) = \log(\frac{P_{ij}}{1-P_{ij}}) = \eta_{ij} = \beta_{0j} + \beta_{qj}.$$

To rearrange equation 5, the probability that person *j* answers item *i* correctly is

$$P_{ij} = \frac{1}{1+exp(-\eta_{ij})}.$$

The second level of MLIRT is the person level, which is essentially the second level of random intercept HLM model with a continuous outcome variable.

In MLIRT, βs are assumed to vary across persons at the item-level and β_{qj} is assumed to be constant across person at the person-level.

Although item parameter βs vary across persons at the item level, when level-1 model and level-2 model are combined, the item parameters β_{qj}, which are β_{1j}, β_{2j} , … , $\beta_{(k-1)j}$, are constant across person and vary across items because there are **no** random terms added to the item effects, i.e., β_{1j} to $\beta_{(k-1)j}$. The person parameter β_{0j}, however, varies across persons and is fixed across items.

Item level of two-level MLIRT

$$\eta_{ij} = \beta_{0j} + \beta_{qj},$$

Person level of two-level MLIRT

$$\beta_{0j} = \gamma_{00} + u_{0j},$$

$$\beta_{1j} = \gamma_{10},$$

$$\beta_{2j} = \gamma_{20},$$

…

$$\beta_{(k-1)j} = \gamma_{(k-1)0}.$$

where u_{0j}, the person ability, is added as a random component of β_{0j} and γ_{00} as a fixed item effect for the reference item.

To combine level-1 and level-2 into one equation,

$$P_{ij} = \frac{1}{1 + exp(-\eta_{ij})}$$

$$= \frac{1}{1 + exp\{-[\mu_{0j} - (-\gamma_{00} - \gamma_{q0})]\}},$$

where $i=q$. The above equation is algebraically equivalent to Rasch model,

$$P_{ij} = \frac{1}{1 + exp[-(\theta_j - (\delta_i)]},$$

where $\theta_j = u_{0j}$, $\delta_i = -\gamma_{00} - \gamma_{q0}$ (Chaimongkol, 2005).

In Rasch IRT model, $\delta_i = -\gamma_{00} - \gamma_{q0}$ are fixed item parameters (constant across persons) and the person parameter θ_j can be considered either fixed or random effects (Kamata, 2001). However, in MLIRT, u_{0j} are random effects, where $u_{0j} \sim N(0, \tau)$.

The **three-level IRT model** adds the school level and estimates the school-level effects. In this framework, the school-level variance of error is explained by the school factors, such as school type (i.e., public school or private school); the student-level variance of error is explained by the student factors, such as gender and SES (Fox, 2005). It is feasible to estimate the effects of the schools on student ability and item difficulty. The ability of a person from a certain school can be divided into two parts, the random effect associated with the school and the average ability of students in the specific school, so that the individual student's ability can also be compared to the school mean ability (Kamata, 2001). The three-level item analysis allows the estimation of school-level abilities as well as person-level abilities, quantifies the variation of person-level random effects across schools, and reveals the interaction effect between the school and the person level. If the school is denoted by *m, m* would be added to the earlier equations and variables in the two-level framework.

Item level of three-level MLIRT

$$\eta_{ijm}=\beta_{0jm}+\beta_{1jm}X_{1ijm}+\beta_{2jm}X_{2ijm}+...+\beta_{(k-1)jm}X_{(k-1)ijm}$$

$$=\log(odds)=\log(\frac{P_{ijm}}{1-P_{ijm}}).$$

Person level of three-level MLIRT

$$\beta_{0jm}=\gamma_{00m}+\mu_{0jm},$$

$$\beta_{1jm}=\gamma_{10m},$$

$$\beta_{2jm}=\gamma_{20m},$$

…

$$\beta_{(k-1)jm} = \gamma_{(k-1)0m},$$

where u_{0jm} is the deviation of the ability of person *j* in school *m* from the mean ability of school *m*, which is γ_{00m}. The variance of u_{0jm} is τ_γ and is assumed to be fixed across schools. γ_{00m} is the overall effect of the reference k^{th} item in school *m* and can be further decomposed into a fixed and random effects. $\gamma_{(k-1)0m}$ is the effect of the $(k-1)^{th}$ item in school *m* ($q=k$-1).

School level of three-level MLIRT

The overall effect of items γ_{00m}, varies across schools and is decomposed into the fixed component π_{000} and the random component r_{00m} at the school level. For school *m*, we have

$$\gamma_{00m} = \pi_{000} + r_{00m},$$

$$\gamma_{10m} = \pi_{100},$$

$$\gamma_{20m} = \pi_{200},$$

…

$$\gamma_{(k-1)0m} = \pi_{(k-1)00},$$

where $\gamma_{00m} \sim N(0, \tau_\pi)$. The item effects, γ_{10m} through $\gamma_{(k-1)0m}$, however, are assumed to be constant across persons. Therefore, no random components are added to π_{100} to $\pi_{(k-1)00}$.

After combine the item-, student-, and school- levels, the finalized model is

$$P_{ijm} = \frac{1}{1 + exp\{-[(\gamma_{00m} + \mu_{0jm}) - (-\pi_{q00} - \pi_{000})]\}},$$

where $-\pi_{q00}-\pi_{000}$ is the item difficulty for item *i* when *i* = *q* (*i*= 1, …, *k*-1) and π_{000} is the item difficulty for the reference item *k* (Kamata, 2001). Compared to the two-level model, the item difficulty terms are similar to the two-level item difficulty, which are $\gamma_{q0}-\gamma_{00}$. The ability terms, however, are slightly different between the two-level and the three-level IRT models. In the three-level model, the abilities for person *j* in school *m* had two components $r_{00m}+u_{0jm}$, in which r_{00m} is the random effect, representing the mean ability of school *m*. u_{0jm} is the person-specific ability of person *j* in school *m*. Therefore, the three-level model is composed of school abilities and person abilities. The ability term in the two-level model, on the other hand, consists only one part, u_{0j}, which is only the person-specific ability of person *j*.

MR-IRT Model

The MR-IRT model combined the two-step MR model to approach the hierarchically structured data and Rasch IRT to estimate the person and item parameters. The multilevel IRT analysis can be performed in three steps.

Step-1: Estimating item and person parameters with Rasch-IRT model

$$P_{ij} = \frac{1}{1+exp[-(\theta_j-\delta_j)]}.$$

Step-2: Regressing the abilities on schools (X_1, X_2, … , X_m)

$$\theta_j=\beta_1 X_1+\beta_2 X_2...+\beta_m X_m+\mu_j,$$

where $X_1 \ldots X_m$ are the dummy variables, representing the school indicators; u_j is the student-specific abilities (random effect).

Step-3: Regressing the estimated coefficients of the school indicators on school-level predictors, if any; otherwise, keeping the **unconditional model**:

$$\beta_m = \pi_m + r_m,$$

where r_m was the school-specific abilities (random effect).

Monte Carlo Simulation

Webster's dictionary defined Monte Carlo simulation as "the use of random sampling techniques and often the use of computer simulation to obtain approximate solutions to mathematical or physical problems especially in terms of a range of values each of which has a calculated probability of being the solution" (Merriam-Webster, Inc., 1994, p. 754-755).With Monte Carlo simulation, one is able to estimate the performance of the theory by replicating the simulations numerous times in order to understand how statistical models behave in real life situations (Davey, Nering & Thompson, 1997; Fan, Felsövályi, Sivo, & Keenan, 2002).

As powerful as Monte Carlo simulation are in replicating situation and hypothetical analysis, simulation studies are not conducted correctly in many situations (Hammersley & Handscomb, 1964). This is because some simulation studies simulated or analyzed datasets under extreme conditions, which are far from reality. Errors, such as measurement error, sampling error, and model specification error, are common in real life and should be considered in simulation design because "simulations were useful only to the extent that they reflect reality" (Davey, Nering & Thompson, 1997, p. 4).

Three real data analyses with Kamata's three-level MLIRT model were conducted by Fox (2004). Table 1 summarizes the conditions of each dataset.

Purposes of the Study

The purposes of the present study were to (a) evaluate the estimation accuracy of person ability and item parameter of Kamata's MLIRT and MR-IRT models in a school setting, (b) investigate the estimates of school-level ability variance of both models and (c) discuss the advantages and disadvantages of each model in accommodating hierarchically structured item response data. This study aimed to compare the performance of the two models for hierarchically structured data under simulation conditions of varying test lengths, sample sizes and intraclass correlation (*ICC*).

Table 1
Parameters of Real Data Sets Used by Fox (2004)

Conditions	Dutch Math Test	Pupils' Performance		West Bank
Number of Schools	72	68		42
Number of Items	18	23	37	50
Student-level Variance	0.767	0.81	0.602	0.408
School-level Variance	0.292	0.19	0.127	0.37
Intraclass Correlation coefficients (ICC)	0.28	0.19	0.17	0.48
Number of Students	2156	3713		3500

Methods

Data Generation

The present simulation study aimed to generate dataset as close to the real data sets that were used by Fox (2004) as possible. Fifty schools were generated; two

conditions of test lengths were simulated: 15 items and 30 items, respectively. In real life, the sample size of students in one school is fairly arbitrary and random. In order to approximate reality, the student sample size was generated as normally distributed in this study, N~ (50, 10) (Figure 1). Because the IRT-*b* is standardized and ranges roughly from -3 to +3 and the extreme easy or difficulty items were not included in the present simulation study. Item difficulty parameters (IRT-*b*) were randomly chosen using a uniform random distribution with values between -2 and 2. To approach the school-level variance statistics of the four real-data studies in Fox (2004) (see Table 1), four conditions of school-level ability variances were created in the present simulation study, which were 0.1, 0.2, 0.3 and 0.4. As stated, the present study was to examine a two-level unconditional model, which meant all the variances should be explained in either the school-level or the student-level. Thus the student-level variances were correspondingly generated as 1 - 0.1 = 0.9, 1 - 0.2 = 0.8, 1 - 0.3 = 0.7 and 1 - 0.4 = 0.6. Intraclass correlation coefficient (*ICC*) was quantified as the school-level variance divided by the total variance, which in the present study, were 0.1 / (0.1 + 0.9) = 0.1, 0.2 / (0.2 + 0.8) = 0.2, 0.3 / (0.3 + 0.7) = 0.3 and 0.4 / (0.4 + 0.6) = 0.4, respectively. The four *ICC* conditions approximated the real *ICC* values in Fox (2004) (see Table 1). The person-specific ability was normally distributed, with a mean of zero and variance of the student-level ability variance. Similarly, the school-specific ability was normally distribution, with a mean of zero and variance of the school-level ability variance. Similarly to Fox (2004), the dichotomous Rasch MLIRT model was used. One hundred replications were generated in this study to ensure more accurate and stable results. This

was chosen because 100 times was the common replication according to the publications of the National Council on Measurement in Education (NCME) in 2009 (Kim, 2010).

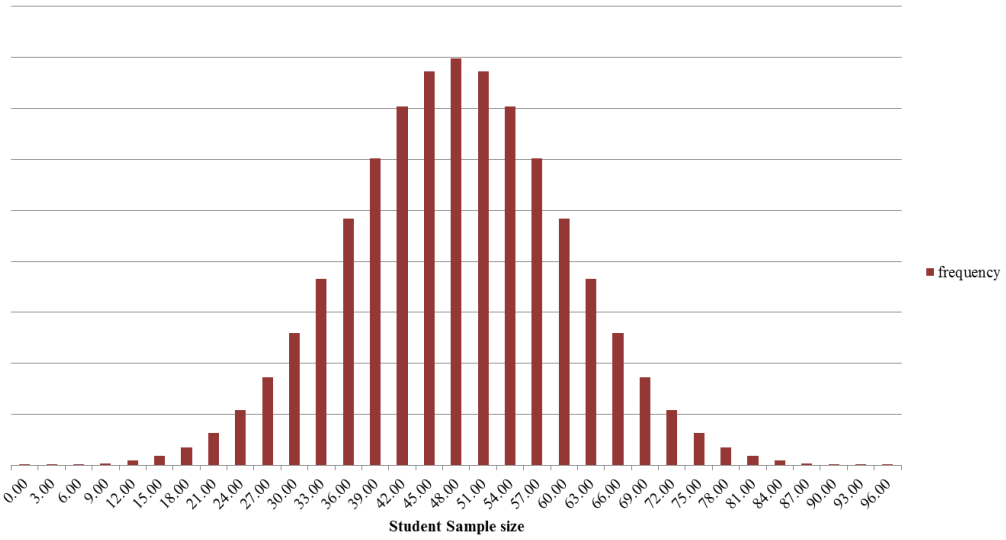

Figure 1. Distribution of student sample size.

Data Analysis

The simulated 1-PL hierarchically structured item response data were generated with SAS 9.3, including the person parameters, item parameters, item responses students from 50 schools. The parameter estimation was performed by Mplus Automation in SAS. The Monte Carlo command in Mplus 7.1 was used to perform 100 replications. The parameter estimation was conducted with both Kamata's MLIRT and MR-IRT models. The estimation of item and person parameters using the MR-IRT model was essentially Confirmatory Factor Analysis (CFA) but with binary factor indicators in the IRT model. The estimation with the MLIRT model would be two-level CFA with binary

factor indicators. The parameter estimation of both models was conducted with Mplus 7.1.

Estimation

The estimation accuracy of abilities and item parameters were assessed by *bias* and the Root Mean Square Error (*RMSE*) and the estimation accuracies of school-level variance were evaluated by the standard error (*SE*) of the variance. *Bias* is defined as the difference between the true value and the average of all possible estimates: $Bias = \eta - E(\eta)$ (Carsey & Harden, 2014). *RMSE* is defined as the square root of the expectation of the estimate squared deviation from the true: $RMSE = \sqrt{E[(\eta - \eta)^2]}$ (Carsey & Harden, 2014). *SE* is an estimate of the variability in a parameter estimate, which gives us an estimate of how certain we are about a given estimate (Carsey & Harden, 2014). *SE* of variance, therefore, helped us to judge how precise the estimate of variance was. Finally, the estimation accuracy of the hierarchically structured item response data by the MR-IRT and MLIRT models was compared.

Results

Item difficulty parameters, person ability and school-level variance were estimated under the conditions of varying test lengths (15 and 30 items), and different *ICC*s (0.1, 0.2, 0.3 and 0.4 respectively). The results were presented in the following five tables. Table 2 and Table3 depicted the estimated accuracy of item difficulty parameters of the two models when there were 15 items and 30 items, respectively, in a test.

The results in Table 2 indicated that MR-IRT and MLIRT both provided accurate item parameter estimates because *bias* and *RMSE* were quite small under all the conditions when using both models. The "Ratio" represented the ratio of mean *bias* (of 100 replications) or *RMSE* of MLIRT over MR-IRT. Because all the ratios were closed to one, there was no significant difference in estimation accuracy of these two models. The *ICC* did not seem to affect the estimation accuracy of item difficulty parameter.

As the number of items increased from 15 to 30, *RMSE* and b*ias* in both models decreased, thus the more accurate the estimate (see Table 3). Again, no considerable difference was found in the estimation accuracy of MLIRT and MR-IRT models because the ratios were closed to one. The increase of *ICC* did not seem to influence the estimation accuracy in either model.

Table 4 and Table 5 presented the estimation accuracy of person ability of the two models when the test length was 15 and 30 respectively. Comparison was conducted based upon *RMSE*, *absolute bias* and *bias*.

Table 4 indicated that, across all *ICC* conditions, the *RMSE*s of person ability estimate were more than twice as much as when using MLIRT model rather than when using MR-IRT model, indicating more accurate ability estimates with a use of MR-IRT model. The values of *bias* of the estimates were about the same when using the two models. The *ICC* did not seem to play an important role in the estimate accuracy of person ability.

Table 5 depicted that, as the number of items increased from 15 to 30, *RMSE* and *bias* of both models decreased. *RMSE*s of person ability estimate were more than

twice as much as when using MLIRT model than when using MR-IRT model. The values of *bias* were quite close of the two models. Therefore, MLIRT was less accurate in estimating person ability than MR-IRT. The change of *ICC* had little influence on the estimate accuracy.

Table 2
The Bias and RMSE of Item Difficulty Parameter of the Two Models When the Test Length was 15

ICC	Statistics	MR-IRT	MLIRT	Ratio
ICC=0.1	**Mean Bias**	-0.0081	-0.0086	1.0601
	RMSE	0.009	0.0095	1.0584
ICC=0.2	**Mean Bias**	0.0066	0.0045	0.6806
	RMSE	0.0081	0.0066	0.8066
ICC=0.3	**Mean Bias**	0.0013	0.0003	0.1975
	RMSE	0.0053	0.0051	0.9668
ICC=0.4	**Mean Bias**	0.0146	0.0146	1.0032
	RMSE	0.0155	0.0155	1.002

Table 3
The Bias and RMSE of Item Difficulty Parameter of the Two Models When the Test Length was 30

ICC	Statistics	MR-IRT	MLIRT	Ratio
ICC=0.1	**Mean Bias**	0.0014	0.0013	0.9404
	RMSE	0.0041	0.0041	0.9956
ICC=0.2	**Mean Bias**	-0.0066	-0.008	1.2103
	RMSE	0.0082	0.0095	1.1631
ICC=0.3	**Mean Bias**	-0.0062	-0.0045	0.7285
	RMSE	0.0081	0.0069	0.8513
ICC=0.4	**Mean Bias**	-0.0049	-0.0042	0.8687
	RMSE	0.0067	0.0063	0.9362

Table 4

The Bias, Absolute Bias and RMSE of Person Ability Estimates of the Two Models When the Test Length was 15

ICC	RMSE						Absolute Bias						Bias					
	MR-IRT			MLIRT			MR-IRT			MLIRT			MR-IRT			MLIRT		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
ICC=0.1	0.1366	0.8289	0.5278 (0.0939)	0.5580	1.7259	1.3210 (0.2517)	1.79E-05	4.7096	0.8012 (0.6134)	1.79E-05	4.7096	0.8012 (0.6156)	-4.7248	4.5414	0.0071 (1.0056)	-4.7248	4.5414	0.0071 (1.0045)
ICC=0.2	0.2029	1.1152	0.5362 (0.0964)	0.3786	1.8971	1.3424 (0.2506)	3.45E-06	4.9920	0.7997 (0.6067)	3.40E-06	4.9000	0.8000 (0.6134)	-4.7115	4.7565	0.0042 (1.0009)	-4.7115	4.7565	0.0042 (1.0005)
ICC=0.3	0.1979	0.7111	0.5328 (0.0942)	0.1998	2.1470	1.3650 (0.2591)	2.10E-05	4.3639	0.8041 (0.6089)	2.10E-05	4.5639	0.8041 (0.6087)	-4.3212	4.3199	0.0035 (1.0047)	-4.3212	4.3199	0.0035 (1.0002)
ICC=0.4	0.2503	0.7592	0.5379 (0.0921)	1.0786	2.6708	1.3866 (0.2628)	9.86E-06	4.5232	0.8105 (0.6326)	9.86E-06	4.6232	0.8105 (0.6400)	-4.8242	4.4525	0.0148 (1.0126)	-4.8242	4.4525	0.0148 (1.0105)

Note. SDs are presented within parentheses.

Table 5

The Bias, Absolute Bias and RMSE of Person Ability Estimates of the Two Models When the Test Length was 30

ICC	RMSE						Absolute Bias						Bias					
	MR-IRT			MLIRT			MR-IRT			MLIRT			MR-IRT			MLIRT		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
ICC=0.1	0.2144	0.5733	0.4074 (0.0713)	0.1529	1.8199	1.2869 (0.2484)	3.54E-05	4.6100	0.7952 (0.6018)	3.54E-05	4.6102	0.7952 (0.6003)	-4.4028	4.5209	0.0024 (0.9998)	-4.4031	4.5207	0.0021 (0.9963)
ICC=0.2	0.0954	0.6065	0.4000 (0.0684)	0.0094	1.8960	1.2911 (0.2449)	9.35E-06	4.5012	0.8102 (0.6074)	9.35E-06	4.5012	0.8106 (0.6136)	-4.3428	4.3967	-0.0076 (1.0175)	-4.3450	4.3946	-0.0098 (1.0163)
ICC=0.3	0.0064	0.3353	0.1692 (0.0292)	0.4411	1.6781	1.2662 (0.2343)	4.46E-06	4.4092	0.7885 (0.6001)	4.46E-06	4.4088	0.7873 (0.6002)	-4.6952	4.4051	-0.0064 (0.9869)	-4.6952	4.4051	-0.0064 (0.9871)
ICC=0.4	0.3264	0.5127	0.4203 (0.0731)	0.6997	2.2437	1.3404 (0.2436)	2.06E-06	4.6993	0.8011 (0.6034)	2.06E-06	4.6992	0.8016 (0.6027)	-4.7635	4.5519	-0.0050 (1.0024)	-4.7632	4.5521	-0.0048 (0.9972)

Note. SDs are presented within parentheses.

Table 6

The Standard Error of Variance of School Ability Variance Estimates

Number of Items	ICC	True School Theta Variance	Estimates of School Theta Variance		Average SE of School Theta Variance		
			MR-IRT	MLIRT	MR-IRT	MLIRT	Ratio
Number of Items=15	ICC=0.1	0.0995	0.0748	0.0961	0.0387	0.0234	0.6050
	ICC=0.2	0.2061	0.0945	0.2008	0.0435	0.0424	0.9754
	ICC=0.3	0.2890	0.1745	0.2820	0.0591	0.0573	0.9701
	ICC=0.4	0.4031	0.2264	0.3925	0.0673	0.0572	0.8500
Number of Items=30	ICC=0.1	0.0999	0.0878	0.0940	0.0419	0.0223	0.5323
	ICC=0.2	0.2045	0.1469	0.1992	0.0542	0.0430	0.7933
	ICC=0.3	0.2956	0.2233	0.2889	0.0668	0.0604	0.9039
	ICC=0.4	0.3956	0.2743	0.3821	0.0741	0.0688	0.9288

In Table 6, the first column presents the true school theta variance across all conditions. The second and third columns present the estimated school theta variances of model MR-IRT and MLIRT, respectively. The fourth and the fifth columns present the *SE* of estimated school theta variances of MR-IRT and MLIRT, respectively. The "Ratio" represents the *SE* of school theta variance of the MLIRT model divided by the *SE* of variance of MR-IRT model.

Across all conditions, the ratios were less than 1, indicating the smaller *SE* of school theta variance of MLIRT model. Thus MLIRT estimated more accurate school ability variances than the MR-IRT model.

The larger the *ICC*, the more variance was explained by the school-level. In other words, the observations were more dependent on each other and more violated of the independency of observations assumption. Therefore, as *ICC* increased, the estimated standard error of school ability variances in both models increased. In either model, across all *ICC* conditions, the *SE* of school ability variances were larger when number of items was 30, which indicated that the longer the test length, the greater was the challenge of estimating the school ability variances.

Discussion

The comparison of MLIRT and MR-IRT models through the present simulation study indicated more accurate performance of MLIRT in estimating the school-level abilities and similar performance of the two models in estimating IRT-*b* parameter. Surprisingly, MLIRT model demonstrated much less accurate estimation for person abilities than MR-IRT model. Two potential explanations were proposed for the less satisfactory performance of MLIRT model in estimating student-level abilities:

(a) MLIRT model assumed the abilities as random effects, which led to generalizability of population parameter, whereas MR-IRT assumed abilities as fixed effects, which resulted in more accurate estimates in sacrifice of generalizability. One of the most critical differences between MLIRT and MR-IRT was to model coefficients as random or fixed (Raudenbush & Bryk, 2002). In the MLIRT model, abilities were assumed to not only vary across students but also across schools. The abilities were treated as random effects at both student level and school level, which meant the student samples and the school samples were randomly selected from the population and can be used to generalize for population parameters. However, the tradeoff was less accurate individual estimates of person abilities. In the MR-IRT model, however, abilities were assumed to be fixed effects, which meant no generalization was made based upon the estimates of the samples. However, when generalizability was not the primary concern, or when sample size was too small for generalization, MR-IRT was more appropriate to use and tended to provide more accurate estimates of person abilities. Such finding provided guidelines for the practitioners in terms of model selection and data collection.

When there are only small samples collected from a few schools in a small school district and the research objective is to accurately evaluate students' abilities, MR-IRT mode is more preferred than MLIRT model. On the other hand, when the research is conducted based upon the hypotheses of large sample sizes from larger areas, (e.g., multiple states or countries), and the goal is to make general evaluation of schools in these areas rather than each individual student, MLIRT is more recommended to serve the research purpose.

(b) Certain extremely small sample sizes in the present study resulted in challenges of generalizability when using MLIRT model. In order to approximate situations in reality, the sample sizes of students in the present simulation study were assumed to be normally distributed, N (50, 10), which took the possibilities of extremely small sample sizes into account. When the sample size was small, it became statistically challenging to generalize the sample to population, which was what MLIRT model intended to do. As a result, in the circumstance of small sample size, MLIRT model had convergence problem in estimating the person abilities, which led to higher overall *RMSE* across the conditions, compared to MR-IRT model. In empirical research, for the purpose of evaluating students' abilities in multiple schools, MLIRT model is recommended when the sample sizes are large across all schools, whereas MR-IRT model is safer to use when small sample sizes exist. The results of the present study were consistent with previous study. Brune (2011) found that when sample size was small, there was very small difference in person- and school- ability estimates between the HGLM (2) (i.e., the two-level IRT model that ignored the random effects among

schools) and HGLM (3) (i.e., the three-level IRT model that considered school- level random effects). Brune (2011) also noted that further analysis should be done before recommending ignoring clustering with small sample sizes, when ability estimates were the primary concern. Brune (2011) recommended a minimum sample size of 100 in order to use 1-P HGLM (i.e., MLIRT model). More specifically, a minimum total sample size of 100 was proposed to accurately estimate level-2 (student-level) residuals and a minimum total sample size of 400 to accurately estimate level-3 (school-level) residuals. Additionally, Brune (2011) concluded the importance of the cluster sample size in estimating student- and school- level abilities.

Limitation and Directions for Future

One area of potential research is the investigation of the impact of sample sizes in parameter estimation. For the purpose of simulating the condition that was closest to reality, the student sample sizes were normally distributed in the present study. The results showed more accurate estimates of person abilities when using MR-IRT than MLIRT. One possible explanation was that MLIRT model failed to generalize the random effects due to some extremely small sample sizes. Brune (2011) suggested that MLIRT would perform better for larger sample size. In future research, multiple conditions of sample sizes should be examined. Comparison of MLIRT and MR-IRT models should be conducted under different sample size conditions to answer the question of when to use which model. Additionally, testing the lower limits of sample size in order to use MLIRT would facilitate practitioners' decision making, in terms of data collection and model selection.

Another limitation of the present study and direction for future research is that the MLIRT model could be expanded for more complicated framework, such as models with predictor variables in each level, 2PL or 3PL dichotomous IRT model or even polytomous models, and multidimensional models, etc. Future research could focus on identifying the optimal conditions (e.g., the optimal sample size and test length) of utilizing MLIRT model in more complex models, in order to obtain accurate estimates of IRT parameters, students' abilities and school accountability.

Conclusions

The present simulation study compared the accuracy of IRT difficulty parameters and abilities estimates of two multilevel IRT models (i.e., MR-IRT and MLIRT). The simulation conditions included (a) test length (15 items and 30 items, and (b) *ICC* (0.1, 0.2, 0.3, and 0.4). The control factor was number of schools, which was held constant as 50 across the 8 conditions. The results indicated that (a) for both MR-IRT and MLIRT models, the longer the test length, the more accurate the estimates of IRT-*b* and person abilities, the less accurate the estimates of the school abilities (i.e., higher *SE*), (b) for both MR-IRT and MLIRT models, *ICC* had very little impact on the estimates of IRT-*b* and person abilities, but high *ICC* resulted in large *SE* of school ability variance estimates, (c) no significant difference was found in estimate accuracy of item difficulty parameter of the two models across all conditions, (d) MLIRT model was much less accurate in estimating person abilities than MR-IRT model across all conditions, and (e) MLIRT model provided more accurate school ability estimates than MR-IRT model across all conditions.

Two potential explanations were given in terms of the less accurate person abilities estimates with MLIRT model. (a) MLIRT focused on parameter generalization by treating abilities as the random effects, whereas compromised the estimating accuracy compared to MR-IRT model. (b) The present simulation included certain extremely small student sample sizes, which resulted in convergence problem when estimating the person abilities with MLIRT model.

In short, when sample sizes are sufficiently large across all schools, MLIRT model is efficient in evaluating school accountability or comparing school abilities, and therefore is preferred when the research interests are at the school level. However, MR-IRT model is more appropriate for evaluating individual student abilities, especially with the existence of small student sample sizes across school.

A LONGITUDINAL MULTILEVEL LOGISTIC REGRESSION MODEL FOR DIF ANALYSES

The presence of differential item functioning (DIF) is a serious problem in educational testing as DIF indicates a threat to the validity of the test (Thissen, Steinberg, & Wainer, 1988, 1993). DIF refers to a differentiation in performance of two groups of examinees with certain different characteristics but same ability level. In other words, different groups of examinees, who are of the same ability, have different probabilities of getting a question correct (Pine, 1977). With the existence of the DIF items, a test may fail to make appropriate inferences or decisions about the examinees' true ability. DIF detection procedures are critical in the validation of educational and psychological tests (Camilli & Shephard, 1994).

The traditional and major DIF detection procedures include, the transformed item difficulty index (Angoff, 1982), analysis of variance (ANOVA) method (Plake, 1981), the Mantel-Haenszel procedure modified by Holland and Thayer (1988), the standardized *p*-difference index of Dorans and Holland (1993), loglinear item response models (Kelderman, 1989), area measures (Raju, 1988), and the non-parametric multidimensional-based IRT approach (Shealy & Stout, 1993a, 1993b). With the traditional methods, the identified DIF items would be reviewed by the test content experts to examine why the items function differently among different subgroups. Corresponding changes would be made to the identified items to address the "biased" estimates (Chaimongkol, 2005).

However, the classic DIF techniques can only detect the existence of DIF, but not explain the sources of DIF. As Kim, Cohen, Alagoz, and Kim (2007) pointed out, the classic DIF detection methods identify the items with DIF and measure the effect sizes of DIF. But little progress has been made in regard to the causes of DIF occurrence (Ferne & Rupp, 2007; Padilla, Prez, & Gonzlez, 1998; Zumbo & Gelin, 2005). In the occasion when there is more than one source of DIF, the classic approaches are not able to interpret the causes of DIF (Roussos & Stout, 1996). Therefore, systematic research needs to be conducted to not only identify DIF but also explain DIF.

One approach to possibly identifying the sources of DIF is the multiple indicator multiple cause (MIMIC) model in the structure equation modeling framework, in which the latent trait is predicted by a group membership variable, in addition to the measurement model. The significance of the path between the individual indicator and the group membership variable represents the occurrence of DIF of a certain item (i.e., the indicator). The major issue with this method in detecting DIF items is that the Type I error rates are high (Finch, 2005; Finch & French, 2011; Wang & Shih, 2010; Woods, 2009; Kim et al., in press). More specifically, Finch and French (2011) found inflated Type I errors and reduced powers in certain conditions of a simulation study when using multilevel MIMIC models for DIF detection. They recommended using multilevel MIMIC models for the purpose of better model fit. Kim, Yoon, Wen, Luo and Kwok (in press) had similar findings when investigating the performance of multilevel MIMIC models in detecting DIF at student-level. Their results indicated high false positive rates (i.e., Type I error rates) for student-level DIF detection.

A reasonable alternative approach of identifying the sources of DIF is the multilevel random effect DIF model (Swanson et al., 2002). The two-level logistic regression model proposed by Swanson et al. (2002) has item responses nested within students to detect DIF and to explain the potential causes of DIF. The level-1 model (item-level) is a logistic regression model for DIF analysis, which is similar to those models proposed by Swanminathan and Rogers (1990). The level-2 model (student-level) treats the coefficients from the level-1 model, including the coefficient(s) that represent DIF, as random effects. The random effects are then predicted by the characteristics of the items in the level-2 model. Therefore, this model can (a) help identify the item characteristics that are related to DIF, (b) estimate the variance of error in DIF that can be explained by these item characteristics and therefore explain the causes of DIF occurrence, and (c) propose alternative causes of the DIF by comparing the explanatory power or fit of different models (Balluerka, Gorostiaga, Gómez-Benito & Hidalgo, 2010).

The mathematical model expression is given as follow.

Level-1 of Multilevel Random Effect DIF Model

$$\text{logit}(P(Y_{ij}=1))=\beta_{0j}+\beta_{1j}*ability_i+\beta_{2j}*group_i,$$

where

Y_{ij} is the item response of person *i* for item *j* (1=correct response, 0=incorrect response),

$ability_i$ represents the ability of person *i* on the certain ability scale of the certain test,

$group_i$ is a grouping dummy variable that represents the subgroup that person *i* belongs to, either a reference group (*group* = 0) or a focal group (*group* = 1),

β_{0j} is the *logit* for the item difficulty in the reference group, which can be interpreted as the item difficulty of reference group,

β_{1j} represents the item discrimination parameter, which has been set as the same value in reference and focal groups in this model,

β_{2j} is the parameter of uniform DIF, which is the difference of item difficulty parameter in the focal group and the reference group.

Of course, for more complex models, the constraint of constant item discrimination and item difficulty parameters between the focal and reference groups can be released, thus enabling both uniform and nonuniform DIF to be modeled. Uniform DIF occurs when the two subgroups only differ in the item difficulty parameter, whereas nonuniform DIF is present when there are disparate differences of the item discrimination parameters and/or pseudo-guessing parameters (Clauser & Mazor, 1998). Additionally, the more complicated models can compare more than two subgroups by using multiple dummy variables to represent the subgroups (Swanson, 2002).

The level-1 coefficients are considered as random effects in the level-2 model and are decomposed to fixed components and random components. The error of variance of the random effects can be predicated by certain item characteristics.

Level-2 of Multilevel Random Effect DIF Model

$$\beta_{0j}=\gamma_{00}+u_{0j},$$
$$\beta_{1j}=\gamma_{10}+u_{1j},$$
$$\beta_{2j}=\gamma_{20}+\gamma_{21}*I_1+\gamma_{22}*I_2+...+\gamma_{2n}*I_n+u_{2j},$$

where

the γ_{k0} s are the grand means of the level-1 coefficients, in which, γ_{00} is the grand mean of the item difficulty parameters and, γ_{10} is the grand mean of the item discrimination parameters in the reference group, γ_{20} is the mean DIF magnitude of item *j*,

γ_{2n} is the coefficient associated with the n^{th} item characteristic in predicting the variation in DIF for item *j*,

the u_{kj}s are random effects with mean zero and variance of 1 that represent unexplained variance. u_{0j} is the variability that cannot be explained by item difficulty parameters, and u_{1j} is the variability that is unexplained by item discrimination parameter, u_{2j} is the variability that cannot be explained by certain item characteristics that result in DIF for item *j*,

$I_1,\ldots, I_n$ are dummy variables, representing item characteristics.

As stated, the Swanson et al. (2002) model contains two levels, student-level nested within item-level. The random effects, DIF across items, are predicted by the item characteristics. However, it is also possible that DIF could be predicted by group unit characteristics, such as schools and communities. In other words, the random effects DIF may also vary across group units. To identify and explain the variation of DIF across group units, Kamata and Binici (2003) modified Swanson's model by including a group-level in a hierarchical linear model (Bryk & Raudenbush, 2002), implemented by the HLM 5 software. Kamata and Binici's model has three levels. Level-1 is the item level, level-2 refers to the student level and level-3 indicates the school level. However,

Kamata and Binici's model is developed based upon the Rasch IRT model and is mathematically challenged to expand to more complex IRT models. Moreover, HLM 5 software package uses the penalized or predictive quasi-likelihood (PQL) method to approximate maximum likelihood (Chaimongkol, 2005) and therefore, produces underestimated level-3 variance of error.

Chaimongkol (2005) proposed a modified multilevel logistic regression approach for DIF analysis on the basis of the Kamata and Binici (2003) method. Similarly, Chaimongkol's (2005) model considered persons, items and group units and modeled a random-effect DIF across group units to explain the sources of DIF by group characteristics variables. However, instead of using HLM5 software for the level-3 variance parameter, Chaimongkol applied Bayesian estimation to obtain more accurate estimates of the level-3 parameters. The disadvantage of this method was that it tended to be quite time consuming.

Recently, a psychometric framework for the evaluation of instructional sensitivity was proposed by Naumann et al. (2013). This framework attempted to interpret DIF that potentially varies across item, group units, and two time points. The changes in item difficulties with interaction of items, time points, and classroom were estimated under a multilevel IRT framework. This model provided an estimate of baseline classroom-specific item difficulty and an estimate of classroom-specific change in item difficulty across two time points. The model then added a latent regression term to yield an explanatory IRT model (Van den Noortgate & De Boeck, 2005) to explain multilevel-DIF and variation of Pretest-Posttest-Difference (PPD) across classes (PPD

variance). However, the model proposed by Naumann et al. (2013) had the following limitations: (a) The model was limited to two time points and was mathematically challenging to extend to more general model with more than two time points, which was neither practical nor applicable in real longitudinal study with more than two time points, (b) the model considered time point as fixed effect, thus no time variance could be estimated, and (c) the theoretical model was applied to the instructional sensitivity analysis with only one empirical data set; no simulation study was performed to evaluate the model under various conditions. Although the application has shown some promising results, without the results from simulation studies, it is quite challenging to identify the most suitable models for different applications (Pastor, 2003).

Literature Review

Many DIF detection procedures have been discussed and suggested conceptually based upon various research objectives (Chaimongkol, 2005). Millsap and Everson (1993) and Potenza and Dorans (1995) wrote good reviews about the DIF detection methods. Angoff (1982) had good insights on the perspectives of DIF methodology and Cole (1993) provided a thorough history and development of DIF methods. The present study only reviewed the logistic regression DIF detection method, which can be expanded to hierarchical logistic regression for the investigation of the random-effect DIF. One-level logistic regression model can be used for simple DIF detection without explaining the causes of DIF. Extended multi-level logistic regression model can not only identify the existence of DIF, but also explain the sources of DIF.

Logistic Regression DIF Detection Method

Logistic regression is used when the outcome variable is dichotomous (0 or 1) whereas the predictor variables are not restricted to certain type. The logistic regression model can be adapted easily to: (a) predicting probabilities of an event, (b) assessing interaction effects of various covariate variables, and (c) understanding the impacts of covariate variables (Garton, 2004).

Study I has demonstrated how IRT models can be represented as logistic regression model (Adams, Wilson, & Wu, 1997; Kamata, 2001). The basic Rasch IRT model (Rasch, 1960) is represented as

$$\text{logit}(P_{ij})=\log\left(\frac{P_{ij}}{1-P_{ij}}\right)=\theta_j-b_i,$$

where $\theta_j \sim N(0, \tau^2)$.

In an IRT framework, logistic regression has two noteworthy advantages: (a) By including person or item covariates as predictor variables, logistic regression models can be used to reformulate the various IRT models and interpret the meanings of the models (De Boeck & Wilson, 2004; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003). (b) With a logistic regression model, the coefficients of logistic regression can be treated as random effects and integrate logistic regression with hierarchical linear model to estimate variance of person and item parameter at higher levels. Such integration results in the multilevel IRT framework, which was discussed in Study I. The hierarchical logistic regression model, which is the multilevel IRT model applied to logistic DIF

detection, is able to explain the causes of DIF at multiple levels by estimating the random variances at each level.

The detection of DIF with logistic regression approach involves statistical significance testing and estimation of DIF magnitude. The logistic regression equation for DIF analysis is given as follows

$$\text{logit}(P(Y_{ij}=1))=\beta_{0j}+\beta_{1j}*ability_{i}+\beta_{2j}*group_{i},$$

where

Y_{ij} is the response of person *j* to item *i* (1=correct response, 0=incorrect response),

$ability_{i}$ represents the ability on a common scale for all students,

$group_{i}$ is dummy variable, where 0 represents the "reference" group and 1 represents the "focal" group,

β_{0j} represents item difficulty parameter of reference group,

β_{1j} represents item discrimination parameter in both reference and focal group (constrained to be equal),

β_{2j} is the magnitude of DIF, which is the deviation of item difficulty from the focal group to the reference group.

To allow item discrimination and item difficulty to differ in reference and focal groups, an alternative logistic regression model for DIF was proposed by Swaminathan and Rogers (1990) by including an interaction term of $ability_{i}$ and $group_{i}$.

$$\text{logit}(P(Y_{ij}=1))=\beta_{0j}+\beta_{1j}*ability_{i}+\beta_{2j}*group_{i}+\beta_{3j}*ability_{i}*group_{i},$$

(Swaminathan & Rogers, 1990).

DIF detection with logistic regression model is conducted by statistically significance test of regression coefficient (Chaimongkol, 2005). If an item is unbiased, only β_{0j} and β_{1j} should be nonzero. Uniform DIF occurs when $\beta_{2j} \neq 0$ and $\beta_{3j} = 0$, while nonuniform DIF occurs when $\beta_{3j} = 0$ (whether or not $\beta_{2j} = 0$) (Swaminathan & Rogers, 1990). Chi-squared difference statistical significance test is later used to detect uniform and nonuniform DIF by (a) fitting the full model with ability, group indicator and interaction of ability and group, (b) removing the interaction effect to obtain the reduced model R_1, (c) removing the group indicator term from R_1 to obtain the reduced model R_2, and (d) calculating the chi-square differences between the full model and R_1, then R_1 and R_2 to detect uniform and nonuniform DIF respectively (Swaminathan & Rogers, 1990). In addition of chi-square difference, Zumbo (1999) proposed computing the differences of the R^2 as the effect size to represent the magnitude of DIF (Chaimongkol, 2005).

Multilevel Logistic Regression

The multilevel logistic regression model is also referred as hierarchical logistic regression model or random effect logistic regression model. The multilevel logistic regression model expands the single-level logistic regression model by considering the coefficient parameters as random effects.

Two-level logistic regression model. As discussed in Study I, the Rasch IRT model is equivalent to the two-level logistic regression model where items are nested within students. The two-level logistic regression model is given as

Level-1 (item-level of two-level logistic regression model)

$$\eta_{ij} = \log(odds) = \log(\frac{P_{ij}}{1-P_{ij}}) = \beta_{0j} + \beta_{1j}X_{1j} + \beta_{2j}X_{2j} + \cdots + \beta_{(k-1)j}X_{(k-1)j},$$

where

P_{ij} is the probability of person *j* answering item *i* correctly and η_{ij} is the item response,

β_{0j} is the intercept, calculated as the predicted *log(odds)* when all X_{ij}= 0,

β_{1j} is the effect associated with item 1, β_{2j} is the effect of item 2, and so on (Natesan, 2007),

X_{qj} is the q^{th} dummy variable for person *j* with a value of 1 when $q = i$ and 0 otherwise (Kamata, 2001).

The effect of item *i* is associated with the q^{th} dummy variable (X_{qj} when $q = i$ and 0 otherwise). Therefore, the level-1 equation is reduced to

$$\log(odds) = \log(\frac{P_{ij}}{1-P_{ij}}) = \eta_{ij} = \beta_{0j} + \beta_{qj}.$$

The mean and variance of item responses η_{ij} are P_{ij} and $P_{ij}(1- P_{ij})$, respectively (Kamata, 2001).

Level-2 (student -level of two-level logistic regression model)

$\beta_{0j} = \gamma_{00} + u_{0j}$,

$\beta_{1j} = \gamma_{10}$,

$\beta_{2j} = \gamma_{20}$,

…

$\beta_{(k-1)j}=\gamma_{(k-1)0}$.

where u_{0j}, the person ability, is added as a random component of β_{0j}, and γ_{00} is a fixed item effect for the reference item.

Three-level logistic regression model. Item responses of students from the same school may have high correlations. To take school-level variance into account, the items are nested within students, who are nested in schools.

If the school is denoted by *m*, *m* would be added to the earlier equations and variables in the two-level logistic regression model.

Level-1 (item-level of three-level logistic regression model)

$$\eta_{ijm}=\beta_{0jm}+\beta_{1jm}X_{1ijm}+\beta_{2jm}X_{2ijm}+...+\beta_{(k-1)jm}X_{(k-1)ijm}$$

$$=\log(odds)=\log(\frac{P_{ijm}}{1-P_{ijm}})=\beta_{0jm}+\beta_{qjm}.$$

Level-2 (student-level of three-level logistic regression model)

$\beta_{0jm}=\gamma_{00m}+u_{0jm}$,

$\beta_{1jm}=\gamma_{10m}$,

$\beta_{2jm}=\gamma_{20m}$,

…

$\beta_{(k-1)jm}=\gamma_{(k-1)0m}$,

where

u_{0jm} is the deviation of person *j* in school *m* from the mean of school *m*, which is γ_{00m},

The variance of u_{0jm} is τ_γ and is assumed to be identical across all schools, γ_{00m} is the overall effect of the reference k^{th} item in school *m* while $\gamma_{(k-1)0m}$ is the effect of the *(k-1)*th item in school *m* ($q = k$-1), assuming the effect the k^{th} item is zero (Kamata, 2001).

Level-3 (school-level of three-level logistic regression model)

The overall effect of item γ_{00m}, is the only term that varies across schools. For school *m*, we have

$$\gamma_{00m} = \pi_{000} + r_{00m},$$

$$\gamma_{10m} = \pi_{100},$$

$$\gamma_{20m} = \pi_{200},$$

…

$$\gamma_{(k-1)0m} = \pi_{(k-1)00},$$

where

$r_{00m} \sim N(0, \tau_\pi)$,

π_{000} is the fixed effect of γ_{00m},

r_{00m} is the random effect of γ_{00m},

τ_π is the variance of γ_{00m},

γ_{10m} through $\gamma_{(k-1)0m}$ are assumed not to vary across person, and therefore only have fixed effects, *i.e.*, π_{100} through $\pi_{(k-1)00}$.

Longitudinal four-level logistic regression model. Higher correlation of the item responses might be found at one time point. The longitudinal logistic regression model includes the time point in the hierarchically structured framework, where items are nested in the time points, which are nested in students who are nested in schools.

If the time point is denoted by *t*, *t* would be added to the earlier equations and variables in the three-level logistic regression model.

Level-1 (item-level of four-level logistic regression model)

$$\eta_{itjm} = \beta_{0tjm} + \beta_{1tjm} X_{1itjm} + \beta_{2tjm} X_{2itjm} + ... + \beta_{(k-1)tjm} X_{(k-1)itjm}$$

$$= \log(odds) = \log(\frac{P_{ijm}}{1 - P_{ijm}}) = \beta_{0tjm} + \beta_{qtjm}.$$

Level-2 (time-level of four-level logistic regression model)

$$\beta_{0tjm} = \gamma_{00jm} + \mu_{0tjm},$$

$$\beta_{1tjm} = \gamma_{10jm},$$

…

$$\beta_{(k-1)tjm} = \gamma_{(k-1)0jm},$$

where

γ_{00jm} is the overall mean effect of the reference k^{th} item of student *j* in school *m* across the time points (fixed effect),

μ_{0tjm} is the deviation of person *j* in school *m* at time point *t* from the mean of student *j* in school *m* across the time points (random effect),

$\gamma_{(k-1)0jm}$ is the mean effect of the $(k-1)^{th}$ item of student *j* in school *m* across the time points (fixed effect).

Level-3 (student-level of four-level logistic regression model)

$$\gamma_{00jm} = r_{000m} + \zeta_{00jm},$$

$$\gamma_{10jm} = r_{100m},$$

…

$\gamma_{(k\text{-}1)0jm} = r_{(k\text{-}1)00m}$,

where

r_{000m} is the mean effect of the reference k^{th} item in school *m* across time points,

ζ_{00jm} is the deviation of person *j* from the overall mean of students in school *m* across time points (random effect),

$r_{(k\text{-}1)00m}$ is the mean effect of the $(k\text{-}1)^{th}$ item of all students in school *m* across all time points (fixed effect).

Level-4 (school-level of four-level logistic regression model)

$r_{000m} = \omega_{0000} + \kappa_{000m}$,

$r_{100m} = \omega_{1000}$,

…

$r_{(k\text{-}1)00m} = \omega_{(k\text{-}1)000}$,

where

ω_{0000} is the grand mean effect of the reference item *k* of all students in all schools across time points,

κ_{000m} is the deviation of school *m* to grand mean effect,

$\omega_{(k\text{-}1)000}$ is the grand mean effect of $(k\text{-}1)^{th}$ item.

Multilevel Logistic Regression Model for DIF Analyses

The multilevel logistic regression DIF model was first outlined by Kamata (2001) who first introduced the multilevel IRT framework (see Study I). The multilevel

logistic regression DIF model is a hybrid of the logistic regression DIF method and multilevel logistic regression IRT model. A Rasch IRT model with DIF parameters can be represented as follows

Level-1 (item-level of multilevel logistic regression for DIF analysis). The first level of logistic regression for DIF analysis model depicted DIF at item level:

$$\text{logit}(P(Y_{ij}=1))=\eta_{ij}=\beta_{0j}+\beta_{1j}X_{1ij}+\beta_{2j}X_{2ij}+...+\beta_{(k-1)j}X_{(k-1)ij}$$

$$=\beta_{0j}+\sum_{q=1}^{k-1}\beta_{qj}X_{qij},$$

where

Y_{ij} is dichotomous item response of person *j* for item *i*, and Y_{ij} is assumed to be independent Bernoulli random variable with the probability of correct response $Y_{ij}=Bernoulli\ (P_{ij})$, $E = P_{ij}$, $Var\ = P_{ij}\ (1-P_{ij})$,

k is the number of items,

for *k* items, only *k*-1 dummy variables X_{qij} are included,

β_{0j} is the level-1 intercept or student main effect, which can be interpreted as the student ability,

β_{qj} stands for the item difficulty for item *q*, which is interpreted relative to the difficulty of the reference item difficulty.

In Kamata's multilevel IRT model, the student abilities vary across students, and therefore are modeled as a random effect at level-2. However, the item difficulty parameters remain constant across students, and are modeled as fixed effect at level-2. When the multilevel IRT model is applied to DIF analysis, the item difficulty parameter

is predicted by the group indicator(s). Therefore, the group indicators are included in the level-2 item difficulty equation.

Level-2 (student-level of multilevel logistic regression for DIF analysis). The second level of logistic regression for DIF analysis model depicted DIF at student level:

$$\beta_{0j}=\gamma_{00}+u_{0j},$$

$$\beta_{qj}=\gamma_{q0}+\gamma_{q1}group_{j},$$

where

γ_{00} is the mean of the student ability, which is set to be equal in focal and reference groups,

u_{0j} is a random effect of β_{0j} and is considered as the specific ability of student *j* from the mean ability, $u_{0j} \sim N(0,\tau_{00})$.

γ_{q0} is the item difficulty parameter of item *q* for the reference group,

γ_{q1} is the difference of item difficulty parameter between the reference and focal group for item *q*, which could also be interpreted as the effect size of uniform DIF for item *q*,

$group_{i}$ is the group indicator that identified the focal group (1) and reference group (0).

To combine the equations in each level together, the Rasch logistic regression DIF model can be represented as

$$\text{logit}(P(Y_{ij}=1))=\gamma_{00}+\sum_{q=1}^{k-1}(\gamma_{qj}+\gamma_{q1}group_{j})X_{qij}+u_{0j}.$$

If the fixed coefficient of an item γ_{q1} is significant, the item is thought to have uniform DIF. In the present simulation study, because a Rasch IRT model was assumed for DIF analysis, the item discrimination parameter was assumed to be equal across groups. Therefore, only uniform DIF could be examined. The multilevel logistic regression model for DIF analysis can be specified and estimated with SAS PROC GLIMMIX (Pan, 2008; Zhu, Rupp & Gao, 2011).

Purposes of the Study

The present study was a Monte Carlo simulation study of longitudinal multilevel DIF approach. The purposes of the present study were to: (a) develop a statistical model that had four nested levels structure in which items were nested within time points, which were nested within students, who were nested in schools, to detect items with DIF, (b) design and conduct simulation study that the DIF of an item may vary across schools and/or multiple time points, (c) identify certain time points, and school characteristics that explain the DIF variation, (d) study how the magnitudes of DIF at the time-level and school-level, the student sample size, the number of schools, the test length, and the percentage of DIF items affected DIF detection results and the variance estimates of the higher levels. To achieve these goals, it was necessary to (a) specify a four-level logistic regression model for DIF analyses with statistical software package SAS 9.3 and (b) assess the adequacy of the proposed models through simulation studies.

Significance of the Study

While the traditional DIF analysis methods can only detect DIF at the item level, the proposed longitudinal multilevel DIF approach identified and explained the causes of

the DIF across time points, students, and/ or schools. Providing this type of information may help in conceptualizing the reasons that items are functioning differentially. The results provided important information to item writers in revising items with DIF, because the fact that the items function differently for a certain type of student, school, or a specific timing can be utilized by the item writers to conceptualize the reasons of DIF occurrence. This approach provided the most general framework to address DIF from the most sources.

Methods

Model Specification

The inherent nested structure of items with DIF is not uncommon in educational testing data. The present study aimed to propose a longitudinal multilevel logistic regression model for DIF analyses to identify and interpret DIF that occurs at multiple levels. This model allowed incorporating (a) time-level characteristics, such as the "beginning of a year" vs. "end of a year" or "pretest vs. posttest", (b) person-level characteristics, such as gender, ethnicity, socio-economic status, and (c) school-level characteristics, such as school types, to describe the composition of the DIF at each level. A general Rasch four-level model for level-1 (items) units *i*, level-2 (time points) units *t*, level-3 (student) units *j*, and level-4 (school) units *m* can be written in a logit form as

Level-1 (DIF associated with item-level characteristics). The first level of the proposed model depicted DIF associated with item indicators:

$$\text{logit}(P(Y_{\text{itjm}}=1)) = \beta_{0\text{tjm}} + \beta_{1\text{tjm}} X_{1\text{itjm}} + \ldots + b_{(k-1)tjm} X_{1\text{itjm}}$$

$$= \beta_{0tjm} + \sum_{q=1}^{k-1} \beta_{qtjm} X_{qitjm},$$

where

X_{qitjm} are dummy variables, representing item indicators,

k is the number of items, and therefore the level-1 model needs *k-1* dummy variables.

Level-2 (DIF associated with time-level characteristics). The second level of the proposed model depicted DIF caused by different time points:

$$\beta_{0tjm} = \gamma_{00jm} + u_{0tjm},$$

$$\beta_{qtjm} = \gamma_{q0jm} + \gamma_{q1jm} * G(time)_j,$$

where

γ_{00jm} is the mean ability of student *j* from school *m* across time points,

u_{0tjm} is the ability deviation of person *j* at school *m* at time *t*, $u_{0tjm} \sim N(0, \tau_1)$,

γ_{q0jm} is item difficulty for item *q* of student *j* in school *m* across time points in the reference group,

γ_{q1jm} is the magnitude of DIF for item *q* across time points,

$G(time)_j$ is the time-level group indicator, with 1 for focal time group and 0 for reference time group.

Level-3 (DIF associated with student-level characteristics). The third level of the proposed model depicted DIF associated with student characteristics:

$$\gamma_{00jm} = r_{000m} + \zeta_{00jm},$$

$$\gamma_{q0jm} = r_{q00m} + r_{q1jm} * G(student)_j,$$

where

r_{000m} is the mean ability of students in school *m* across time points,

ζ_{0q0jm} is the student ability for student *j* deviated from the mean r_{000m}, $\zeta_{00jm} \sim N(0, \tau_2)$,

r_{q00m} is the item difficulty parameter for item *q* of all students in school *m* in the reference group,

r_{q1jm} is the difference of item difficulty parameter between the student reference and focal group for item *q*, which could also interpret as the effect size of uniform DIF for item *q* due to student characteristics,

$G(student)_j$ is the student-level group indicator, with 1 for focal student group and 0 for reference student group.

Level-4 (DIF associated with school-level characteristics). The fourth level of the proposed model depicted DIF associated with school characteristics:

$$r_{000m} = \omega_{0000} + \kappa_{000m},$$

$$r_{q00m} = \omega_{q000} + \omega_{q10m} * G(school)_j,$$

where

ω_{0000} is the grand mean ability of all students from all schools across time,

κ_{000m} is the student ability for students from school *m* deviated from the grand mean ω_{0000}, $\kappa_{000m} \sim N(0, \tau_3)$,

ω_{q000} is the item difficulty parameter for item *q* of all students from all schools across time in the reference group,

ω_{q10m} is the difference of item difficulty parameter between the school reference and focal group for item *q*, which could also be interpreted as the effect size of uniform DIF for item *q* due to school characteristics,

$G(school)_j$ is the school-level group indicator, with 1 for focal school group and 0 for reference school group.

There are two things that are worth mentioning:

(a) All the random effects from each level u_{0tjm}, ζ_{00jm} and κ_{000m} were independent from each other and followed multivariate normal distribution.

(b) There could be more than one set of focal and reference groups at a certain level if there were multiple characteristics that contributed to the DIF at that level. Each set would have their own equations for DIF detection. The significance of DIF caused by the different characteristics can be tested with the equations associated with the certain characteristics. For example, in the time-level, if there were 4 time points and we were interested in finding which time-point contributed to the detected DIF, we would have four sets of equations in the time-level:

$$\beta_{0tjm} = \gamma_{00jm(1)} + u_{0tjm(1)},$$

$$\beta_{qtjm} = \gamma_{q0jm(1)} + \gamma_{q1jm(1)} * G(time1)_j,$$

$$\beta_{0tjm} = \gamma_{00jm(2)} + u_{0tjm(2)},$$

$$\beta_{qtjm} = \gamma_{q0jm(2)} + \gamma_{q1jm(2)} * G(time2)_j,$$

$$\beta_{0tjm} = \gamma_{00jm(3)} + u_{0tjm(3)},$$

$$\beta_{qtjm} = \gamma_{q0jm(3)} + \gamma_{q1jm(3)} * G(time3)_j,$$

$$\beta_{0tjm} = \gamma_{00jm(4)} + u_{0tjm(4)},$$

$$\beta_{qtjm} = \gamma_{q0jm(4)} + \gamma_{q1jm(4)} * G(time4)_j.$$

In contrast to the traditional DIF procedures, the longitudinal multilevel DIF model can not only detect the existence of DIF but also the causes of DIF occurrence. The current DIF procedures identify DIF by taking the group characteristics that were associated with DIF into account, such as gender, race and ethnicity, but they could not explain how each examinee was affected differentially and why examinees answered items differentially (Atar, 2006; Cohen, Gregg, & Deng, 2004). The longitudinal multilevel DIF model in the present study, however, included the differentiating factors that may vary among students, time points, and schools as predictor variables at multiple levels of the model to explain the causes of DIF occurrence.

Simulation Design

A Monte Carlo simulation study was designed to (a) examine the performance of the multilevel logistic regression model for detecting DIF, (b) study how the magnitudes of DIF at time- and school- levels, the proportion of items with DIF, the test length, the student sample size and school size affect DIF analysis, and (c) study the Type I error rates and Power of DIF detection when DIF occurs at the school level, the time level and at both levels. The specific research questions addressed in the present study were as follows:

(a) Does the proposed model correctly detect the time-level and/or school-level DIF with the time-level and school-level covariates?

(b) How would the magnitudes of DIF at both levels affect the detection of DIF occurring at both levels?

(c) How would the percentage of DIF items at both levels affect the detection of DIF occurring at both levels?

(d) How would the student sample size affect the detection of DIF occurring at both levels?

(e) How would the test length affect the detection of DIF occurring at both levels?

The study was designed on the basis of the study by Narayanan and Swaminathan (1996) with some modifications on the conditions to serve the objectives of the present study better. Simulated conditions of the present study included the magnitudes of DIF at time level and school level, student sample size, test length, and the proportion of items with DIF at school level and at time level. Other factors that were manipulated in the simulation study included the number of schools, the number of time points, and the proportion of reference and focal groups at each level. Simulated data generation and parameter estimates were conducted with SAS 9.3.

The Rasch IRT model was applied to generate the simulated data. The item difficulty parameter *b* was randomly chosen using a random normal distribution with values between -2 and 2, because the IRT-*b* is standardized and ranges roughly from -3 to +3 and the extremely easy or difficulty items were not included in the present

simulation study. The item discrimination parameter was set to be constant 1 to simplify the model. The abilities of the schools were generated at random from a standard normal distribution, N (0, 1), and abilities of students were generated from a multivariate normal distribution with mean of school-level abilities and variance of an identity matrix. This was the same as the condition used by Wen (2014), which ensured that there was variability at the student level that could be explained by the school level. The random effects at the time-level were generated independently from the student-level and the school-level random effects, following a Toeplitz 1 correlation matrix structure. The correlation matrix had 1s on the main diagonal and the population pretest-posttest correlation coefficient of 0.81, which was reported by Cole et al. (2011) with the use of population achievement data from four states and two large districts in English/Language Arts and Mathematics for the recent three years. The means were assumed to be 0.

Item difficulty parameters were generated, following a standard normal distribution, which was proposed by Wen (2014). Time-level DIF items were generated by adding the magnitude of DIF_t (0.5 or 1) to the difficulty parameters of the focal group (i.e., the pretest) while the reference group (i.e., the posttest) remain unchanged. At the time-level, DIF was simulated based on the rationale that a students' performance on particular items would be influenced by when the test was taken (i.e., either a pretest or a posttest). The pretest (i.e., the focal group) was coded as 1 while the posttest (i.e., the reference group) was coded as 0. When DIF only occurred at the school-level, the difficulty parameter for reference group examinees was generated from $b \sim N$ (0, 1),

whereas the difficulty parameter for students in the focal group was modified as b + DIF_{sch}, which in the present study, DIF_{sch}=0.5 or 1. At the school-level, DIF was simulated based on the rationale that a students' performance on particular items would be influenced their schools' mean socioeconomic status (SES) (i.e., low SES *vs.* high SES). The low SES group (i.e., the focal group) was coded as 1 while the high SES group (i.e., the reference group) was coded as 0.When DIF was presented at both the school and time levels, items that functioned differentially at the time level were more difficult for students at the pretest, and were unchanged for students at the posttest; items that functioned differently at the school level were more difficult for the low SES group than for the high SES group. Table 7 depicted the way how multilevel DIF was generated at each level (Wen, 2014).

Table 7
Generating Multilevel DIF Items

	Manifest groups			
	High SES school		**Low SES school**	
	Posttest	**Pretest**	**Posttest**	**Pretest**
School-level only	*b*		*b+DIFsch*	
School-time levels	*b*	*b+DIFt*	*b+DIFsch*	*b+DIFsch+DIFt*

Two test lengths were selected to investigate the performance of the proposed DIF procedure, 20 items and 40 items. The test lengths were selected because most of the scales and questionnaires in psychological assessment include between 20 and 40 items (Balluerka, Gorostiaga, Gómez-Benito & Hidalgo, 2010; Whitmore &

Schumacker, 1999). For instance, the State-trait Anxiety Inventory (STAI) had 40 items and a STAI for children (STAIC) had the same number of items (Julian, 2011). There were 39 items on the Writing Skills section of the PSAT/NMSQT test and 38 items on the Mathematics section of the PSAT/NMSQT test (Cho, 2007). The Beck Anxiety Inventory (BAI) had a total of 21 items (Julian, 2011).

Another factor changed was the percentage of items with DIF. Ten percent of DIF items and 30 percent of DIF were used at the time- and school- levels. The percentages of DIF items were chosen based on previous research which suggested that it is common to observe between 10% and 30% of items functioning differentially on many tests (Hambleton & Rogers, 1989; Raju, Bode, & Larsen, 1989).

Different student sample sizes were generated to investigate the impact of student sample size on the DIF detection. Kim (2010) pointed out that "…sample size was a core factor for detecting DIF… A small sample size could cause a poor estimation, resulting in true DIF items not being detected… A large sample size could result in precise detection of true DIF items, although the possibility exists that items with no DIF will be detected as if they are true DIF items" (p. 19). Because the sample size ratio between the reference and focal group varied widely across previous simulation studies, the percentage, 50%, was arbitrarily set for convenience here (Pan, 2008). In most DIF research, the range of sample sizes was between 500 and 5000 (Kim, 2010). A real data set example is Spring 2004 Florida Comprehensive Assessment Test (FCAT) science data for 10^{th} grade students used by Atar (2006) for DIF analysis. The sample sizes of the three subsets are 600, 1200, and 2400. Therefore, three similar sample size

conditions were simulated in the present study: (a) 500 total (200 in each group), (b) 1000 total (500 in each group), and (c) 2000 total (1000 in each group).

The school sizes were simulated based on the real data set from The College Board SOAS report, which had 25 students per school participating in the PSAT/NMSQT program. The total students sample sizes yielded the school sizes to be 500 / 25 = 20, 1000 / 25 = 40, and 2000 / 25 = 80, respectively (Cho, 2007).

The DIF magnitude was simulated by varying the item difficulty parameter by 0.5 and 1 and by holding the item discrimination parameters constant as one for the reference and focal groups. DIF was generated as the difference between the item difficulty parameters across the school groups or different time points. Pan (2008) showed that the DIF value of 0.6 was large enough for the multilevel logistic regression model to find most of DIF items. The amount of 0.9, which was used by Hidalgo and Lopez-Pina (2004), was large enough to find the DIF items under most conditions but also tended to have higher Type I error rates. The DIF were generated to favor the reference group over the focal group, which meant that the items with DIF were more difficult for the focal group than they are for the reference group by 0.5 or 1 on the logit scale (Pan, 2008).

To sum up, the following conditions were examined in the present simulation study: (a) test lengths (20 and 40 items), (b) the percentage of items with DIF at school-level (10% and 30%), (c) the percentage of items with DIF at time-level (10% and 30%), (d) student sample sizes (500, 1000, and 2000), (e) the magnitude of school-level DIF (0.5 and 1) (f) the magnitude of time-level DIF (0.5 and 1). There was a total of 2 * 2 * 3

* 2 * 2 = 48 conditions simulated for the present study. One hundred replications were generated to ensure more accurate and stable results. This was chosen because 100 times was the common replication according to the publications of the National Council on Measurement in Education (NCME) in 2009 (Kim, 2010).

Estimation

Logistic regression models estimated by Maximum Likelihood (ML) can be investigated using SAS PROC LOGISTIC. The multilevel logistic regression model, although more complicated to estimate by ML in SAS, was still manageable. The relevant approaches can mainly be classified into two broad categories, linearization and integral approximation methods (Schabenberger & Pierce, 2002). Pan (2008) thoroughly reviewed on which estimation approach should be used in multilevel logistic regression model, specifically when conducting research in SAS. The linearization method was concluded to be the better estimation method for multilevel logistic regression model because it considers complex R-side covariate structures. In the present study, SAS PROC GLIMMIX was used for estimation because it implements one linearization-based method—Pseudo-likelihood (Pan, 2008; SAS Institute Inc., 2011).

The accuracy of DIF detection of time level and school level was evaluated under each condition by power and Type I error rates for detecting uniform DIF. Power was defined as the probability that an item that had DIF was identified, which was calculated by the proportion of times that DIF was correctly identified or the proportion of cases in which DIF items were correctly detected. Type I error rate was the probability that an item was identified to have DIF which in fact, did not have it, which was calculated as

the proportion of times that a non-DIF item was falsely identified (Kristjansson et al., 2005; Pan, 2008).

Results

Type I error refers to the false detection of an invariant item as non-invariant (Wen, 2014) and is based on exact distribution assuming the procedures adhere well to the nominal level of alpha (Nandakumar & Rossos, 2011). In this study, Type I error rates was evaluated at the 0.05 level. According to Bradley (1978), the acceptable range of Type I error rates is computed with a formulae $\alpha \pm ½\alpha$. When $\alpha = 0.05$, the Type I error rates between .025 and .075 are considered reasonable.

The test-wide Type I error rate for the present study was calculated in the following way. Firstly, for each replication, the occurrences of non-DIF items being falsely identified as DIF items were counted. Secondly, the proportion of these false positive counts was calculated for each replication, i.e., in each test. Type I error rate of a test was the average of these proportions (Kim, 2010). Different conditions were examined and the Type I error rates and power of each condition were reported in two tables. Across all the 48 conditions, Type I error rates ranged from 0.021 to 0.089 at the time-level and 0 to 0.153 at the school level, which both roughly fell in the Bradley's (1978) liberal robustness criterion range of 0.025 to 0.075 (see Table 8). The time-level highest Type I error (0.089) occurred when DIF magnitudes at both levels were 0.5, the number of school was the smallest (20), test length was the shortest (20) and percentage of DIF items were smaller (0.1). The time-level lowest Type I error (0.021) occurred under two conditions: (a) when time-level DIF magnitude=1, school-level DIF=0.5, the

sample size was 20, test length was 20 and the percentage of DIF items was 0.3, (b) when DIF magnitudes at both levels were 1, sample size was 40, test length was 20 and the percentage of items with DIF was 0.3.

Power refers to the true detection of an invariant item and was used to investigate how well the longitudinal multilevel logistic regression model performed in terms of DIF detection. Power is defined as the proportion of cases in which DIF items are correctly detected. Any value that is equal or larger than 0.8 is presumed to be indicative of high power (Wen, 2014). Across all the 48 conditions, power ranged from 0.8 to 0.992 at the time level while 0.133 to 0.992 at the school level. Overall, DIF analysis at the time level yielded almost perfect results (power$\geq$ 0.8). On the contrary, power of the DIF detection at the school level was much lower (see Table 9). The majority of the conditions yielded high power of 0.992 at both levels. The lowest power at time level (0.8) occurred when the DIF magnitudes of both levels were 0.5, number of schools was 20, test length was 20 and percentage of DIF items was 0.1. The worst condition power at school level was only 0.133. This occurred when the magnitude of student-level DIF was 1, the magnitude of time-level DIF was 0.5, the sample size was 20, the test length was 20 and the percentage of DIF items was 0.3.

Table 8

Type I Error Rates

Conditions	Time-level DIF magnitude	School-level DIF magnitude	Number of schools	Test length	Percentage of DIF items	Type I error rate(time)	Type I error rate(school)
1	0.5	0.5	20	20	0.1	0.089	0.017
2	0.5	0.5	20	20	0.3	0.071	0.071
3	0.5	0.5	20	40	0.1	0.086	0.044
4	0.5	0.5	20	40	0.3	0.043	0.089
5	0.5	0.5	40	20	0.1	0.057	0.011
6	0.5	0.5	40	20	0.3	0.029	0.036
7	0.5	0.5	40	40	0.1	0.058	0.044
8	0.5	0.5	40	40	0.3	0.061	0.107
9	0.5	0.5	80	20	0.1	0.033	0.011
10	0.5	0.5	80	20	0.3	0.029	0.014
11	0.5	0.5	80	40	0.1	0.056	0.028
12	0.5	0.5	80	40	0.3	0.046	0.014
13	0.5	1	20	20	0.1	0.039	0.022
14	0.5	1	20	20	0.3	0.036	0.029
15	0.5	1	20	40	0.1	0.036	0.153
16	0.5	1	20	40	0.3	0.050	0.000
17	0.5	1	40	20	0.1	0.056	0.011
18	0.5	1	40	20	0.3	0.036	0.071
19	0.5	1	40	40	0.1	0.058	0.064
20	0.5	1	40	40	0.3	0.032	0.014
21	0.5	1	80	20	0.1	0.033	0.044
22	0.5	1	80	20	0.3	0.057	0.050
23	0.5	1	80	40	0.1	0.042	0.072
24	0.5	1	80	40	0.3	0.046	0.057
25	1	0.5	20	20	0.1	0.039	0.000
26	1	0.5	20	20	0.3	0.021	0.050
27	1	0.5	20	40	0.1	0.044	0.028
28	1	0.5	20	40	0.3	0.032	0.025
29	1	0.5	40	20	0.1	0.028	0.006
30	1	0.5	40	20	0.3	0.036	0.036
31	1	0.5	40	40	0.1	0.042	0.022
32	1	0.5	40	40	0.3	0.036	0.057
33	1	0.5	80	20	0.1	0.056	0.011
34	1	0.5	80	20	0.3	0.057	0.036
35	1	0.5	80	40	0.1	0.047	0.025
36	1	0.5	80	40	0.3	0.050	0.032
37	1	1	20	20	0.1	0.028	0.056
38	1	1	20	20	0.3	0.043	0.086
39	1	1	20	40	0.1	0.061	0.042
40	1	1	20	40	0.3	0.061	0.036
41	1	1	40	20	0.1	0.078	0.100
42	1	1	40	20	0.3	0.021	0.064
43	1	1	40	40	0.1	0.050	0.014
44	1	1	40	40	0.3	0.050	0.011
45	1	1	80	20	0.1	0.050	0.061
46	1	1	80	20	0.3	0.043	0.050
47	1	1	80	40	0.1	0.061	0.039
48	1	1	80	40	0.3	0.036	0.029

Table 9
Power

Conditions	Time-level DIF magnitude	School-level DIF magnitude	Number of schools	Test length	Percentage of DIF items	Power(time)	Power(school)
1	0.5	0.5	20	20	0.1	0.800	0.600
2	0.5	0.5	20	20	0.3	0.867	0.550
3	0.5	0.5	20	40	0.1	0.850	0.250
4	0.5	0.5	20	40	0.3	0.867	0.558
5	0.5	0.5	40	20	0.1	0.992	0.900
6	0.5	0.5	40	20	0.3	0.992	0.700
7	0.5	0.5	40	40	0.1	0.975	0.675
8	0.5	0.5	40	40	0.3	0.975	0.558
9	0.5	0.5	80	20	0.1	0.992	0.950
10	0.5	0.5	80	20	0.3	0.992	0.967
11	0.5	0.5	80	40	0.1	0.975	0.950
12	0.5	0.5	80	40	0.3	0.992	0.967
13	0.5	1	20	20	0.1	0.850	0.992
14	0.5	1	20	20	0.3	0.850	0.900
15	0.5	1	20	40	0.1	0.825	0.900
16	0.5	1	20	40	0.3	0.825	0.992
17	0.5	1	40	20	0.1	0.950	0.992
18	0.5	1	40	20	0.3	0.983	0.992
19	0.5	1	40	40	0.1	0.975	0.992
20	0.5	1	40	40	0.3	0.958	0.992
21	0.5	1	80	20	0.1	0.992	0.992
22	0.5	1	80	20	0.3	0.992	0.992
23	0.5	1	80	40	0.1	0.992	0.992
24	0.5	1	80	40	0.3	0.992	0.992
25	1	0.5	20	20	0.1	0.992	0.500
26	1	0.5	20	20	0.3	0.992	0.133
27	1	0.5	20	40	0.1	0.992	0.575
28	1	0.5	20	40	0.3	0.992	0.450
29	1	0.5	40	20	0.1	0.992	0.750
30	1	0.5	40	20	0.3	0.992	0.883
31	1	0.5	40	40	0.1	0.992	0.800
32	1	0.5	40	40	0.3	0.992	0.683
33	1	0.5	80	20	0.1	0.992	0.950
34	1	0.5	80	20	0.3	0.992	0.967
35	1	0.5	80	40	0.1	0.992	0.992
36	1	0.5	80	40	0.3	0.992	0.942
37	1	1	20	20	0.1	0.992	0.900
38	1	1	20	20	0.3	0.983	0.917
39	1	1	20	40	0.1	0.992	0.925
40	1	1	20	40	0.3	0.992	0.983
41	1	1	40	20	0.1	0.992	0.992
42	1	1	40	20	0.3	0.992	0.992
43	1	1	40	40	0.1	0.992	0.992
44	1	1	40	40	0.3	0.992	0.992
45	1	1	80	20	0.1	0.992	0.992
46	1	1	80	20	0.3	0.992	0.992
47	1	1	80	40	0.1	0.992	0.992
48	1	1	80	40	0.3	0.992	0.992

In order to evaluate how well the multilevel logistic regression model performed in DIF detection at different levels, one-way between subjects univariate Analyses of variances (ANOVAs) were conducted on Type I error rate and power. The impact of each factor on Type I error rates and power of DIF detection was summarized in Table 10 to Table 11, in which the factors that were statistically significant and were associated with relatively large effect size (η^2>0.05) were marked as bold.

Table 10
Statistical Significance (p-value) of Each Condition in One-way ANOVA Analysis

	Factors				
Statistical significance (*p*)	**Time-level DIF magnitude**	**School-Level DIF magnitude**	**Number of schools**	**Test length**	**Percentage of DIF items**
Power (time)	**0.000**	0.816	**0.000**	0.915	0.785
Type I error rate (time)	0.304	0.687	0.827	0.257	0.052
Power (school)	0.969	**0.000**	**0.000**	0.812	0.760
Type I error rate (school)	0.453	0.093	0.620	0.633	0.520

Table 11
Effect Size (η^2) of Each Condition in One-way ANOVA Analysis

	Factors				
Effect size (η2)	**Time-level DIF magnitude**	**School-Level DIF magnitude**	**Number of schools**	**Test length**	**Percentage of DIF items**
Power (time)	**0.252**	0.001	**0.361**	0.000	0.002
Type I error rate (time)	0.023	0.004	0.008	0.028	**0.080**
Power (school)	0.000	**0.364**	**0.299**	0.001	0.002
Type I error rate (school)	0.012	**0.060**	0.021	0.031	0.009

The influence of each condition of each factor (i.e. time-level DIF magnitude, school-level DIF magnitude, sample size, test length, and percentage of DIF items) on power and Type I error rates were summarized in Table 12 to Table 15, which indicated how each level of a certain factor differed from each other. The magnitude of time- and

school- level DIF, and the sample size were found to be influential in impacting the power.

The ANOVA analyses indicated that (a) the time-level DIF magnitude had a statistically significant impact on power of DIF detection at the time-level, p= 0.000 and had a large effect size (η^2) of 0.252, (b) the school-level DIF magnitude had a statistically significant impact on power of DIF detection at the school-level, p= 0.000 and had a large effect size (η^2) of 0.364, (c) sample size had statistically significant effects on both the time- and school- level power, with both p= 0.000 and the corresponding effect sizes (η^2) = 0.361 and 0.299, respectively. No statistical significance was found in factors of "test length" and "percentage of DIF items". However, the factor "percentage of DIF items" was found have a relatively large effect (η^2>0.05) on Type I error rate at the time level, η^2= 0.08 and the corresponding p-value was 0.052. Effect size of the school-level DIF magnitude on the school-level Type I error rate was also fairly large, η^2= .06.

Table 12
The ANOVA of Time-level Type I Error Rates

Source	Tests of Within-Sujects Contrasts			Tests of Between-Subjects Effects					
	Sum of Squares	df	Mean Square	Sum of Squares	df	Mean Square	F	p	η2
Time-level DIF magnitude	0.011	46	0.000	0.000	1	0.000	1.080	0.304	0.023
School-level DIF magnitude	0.011	46	0.000	0.000	1	0.000	0.164	0.687	0.004
Number of schools	0.011	45	0.000	0.000	2	0.000	0.190	0.827	0.008
Test length	0.011	46	0.000	0.000	1	0.000	1.316	0.257	0.028
Percetage of DIF items	0.010	46	0.000	0.001	1	0.001	3.994	0.052	**0.080**

Table 13
The ANOVA of School-level Type I Error Rates

Source	Tests of Within-Sujects Contrasts			Tests of Between-Subjects Effects					
	Sum of Squares	df	Mean Square	Sum of Squares	df	Mean Square	F	p	η2
Time-level DIF magnitude	0.044	46	0.001	0.001	1	0.001	0.573	0.453	0.012
School-level DIF magnitude	0.042	46	0.001	0.003	1	0.003	2.947	0.093	**0.060**
Number of schools	0.044	45	0.001	0.001	2	0.000	0.483	0.620	0.021
Test length	0.044	46	0.001	0.000	1	0.000	0.232	0.633	0.005
Percetage of DIF items	0.044	46	0.001	0.000	1	0.000	0.420	0.520	0.009

Table 14
The ANOVA of Time-level Power

Source	Tests of Within-Sujects Contrasts			Tests of Between-Subjects Effects					
	Sum of Squares	df	Mean Square	Sum of Squares	df	Mean Square	F	p	η2
Time-level DIF magnitude	0.112	46	0.002	0.038	1	0.038	15.475	**0.000**	**0.252**
School-level DIF magnitude	0.150	46	0.003	0.000	1	0.000	0.055	0.816	0.001
Number of schools	0.096	45	0.002	0.054	2	0.027	12.702	**0.000**	**0.361**
Test length	0.150	46	0.003	0.000	1	0.000	0.011	0.915	0.000
Percetage of DIF items	0.150	46	0.003	0.000	1	0.000	0.075	0.785	0.002

Table 15
The ANOVA of School-level Power

Source	Tests of Within-Sujects Contrasts			Tests of Between-Subjects Effects					
	Sum of Squares	df	Mean Square	Sum of Squares	df	Mean Square	F	p	η2
Time-level DIF magnitude	2.148	46	0.047	0.000	1	0.000	0.002	0.969	0.000
School-level DIF magnitude	1.365	46	0.030	0.783	1	0.783	26.376	**0.000**	**0.364**
Number of schools	1.506	45	0.033	0.642	2	0.321	9.593	**0.000**	**0.299**
Test length	2.146	46	0.047	0.003	1	0.003	0.057	0.812	0.001
Percetage of DIF items	2.144	46	0.047	0.004	1	0.004	0.094	0.760	0.002

In order to further understand the impacts of each condition on power and Type I error rates at time and school levels, (a) Figure 2 to Figure 5 depicted the effects of time-level DIF magnitude on the mean power and mean Type I error rate at both levels, (b) Figure 6 to Figure 9 depicted the effects of school-level DIF magnitude on the mean power and mean Type I error rate at both levels, (c) Figure 10 to Figure 13 depicted the effects of sample size on the mean power and mean Type I error rate at both levels, (d) Figure 14 to Figure 17 depicted the effects of test length on the mean power and mean Type I error rate at both levels, and (e) Figure 18 to Figure 21 depicted the effects of percentage of DIF items on the mean power and mean Type I error rate at both levels. Only the significant impacts were interpreted in the present work.

As demonstrated in Figure 2, time-level power increased as DIF magnitude increased. When time-level DIF magnitude= 0.5, power was 0.936; whereas when school-level DIF magnitude=1, power was 0.992.

Figure 6 showed that power increased as the school-level DIF magnitude increased as when school-level DIF=0.5, power was 0.719, whereas when school-level DIF=1, power was 0.974. Figure 8 depicted that the increase of school-level DIF magnitude led to an increase of school-level Type I error, with Type I error rate being 0.034 when DIF=0.5 and .0049 when DIF=1.

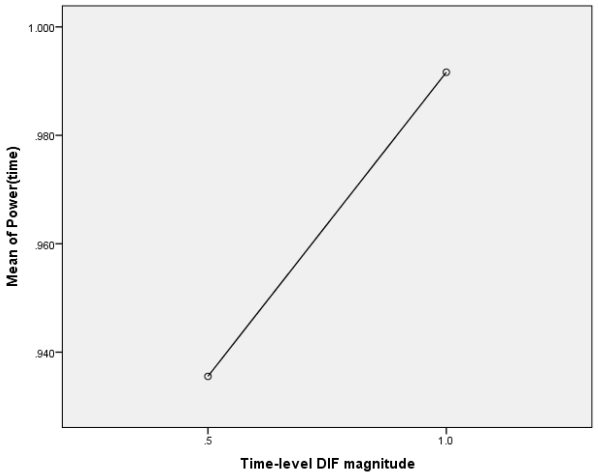

Figure 2. Impact of time-level DIF on time-level power.

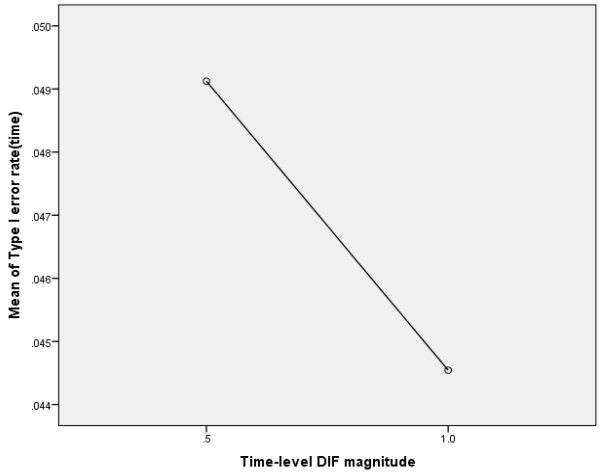

Figure 3. Impact of time-level DIF on time-level Type I error.

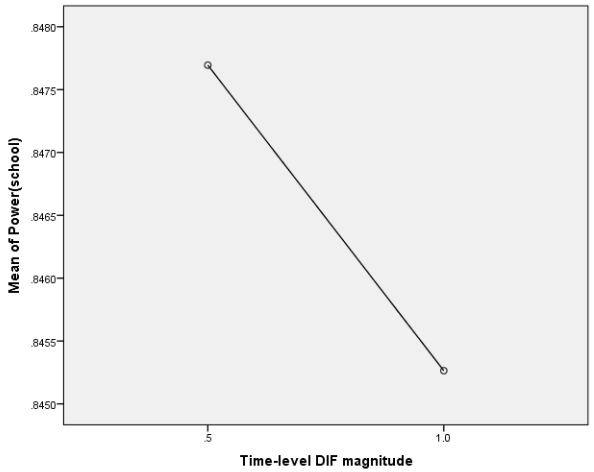

Figure 4. Impact of time-level DIF on school-level power.

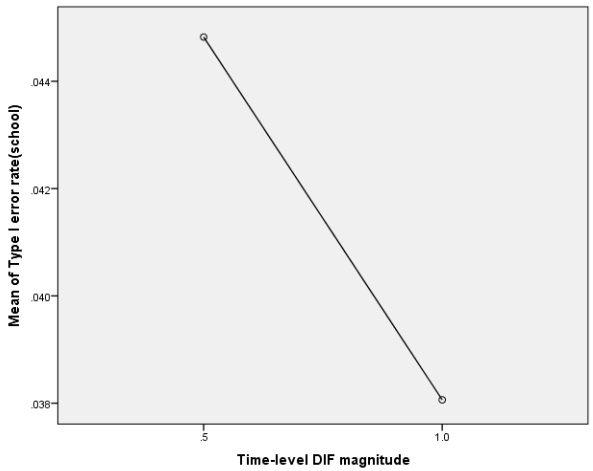

Figure 5. Impact of time-level DIF on school-level Type I error.

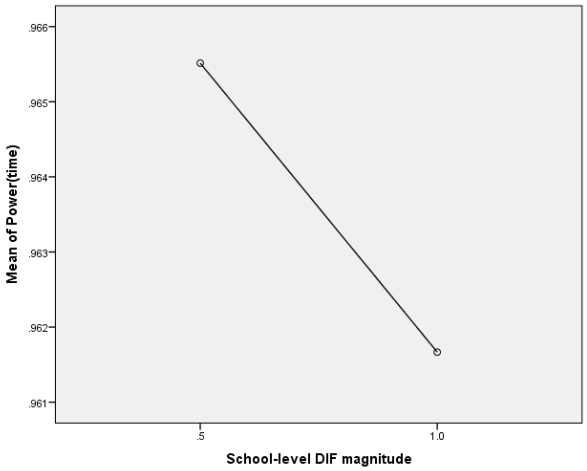

Figure 6. Impact of school-level DIF on time-level power.

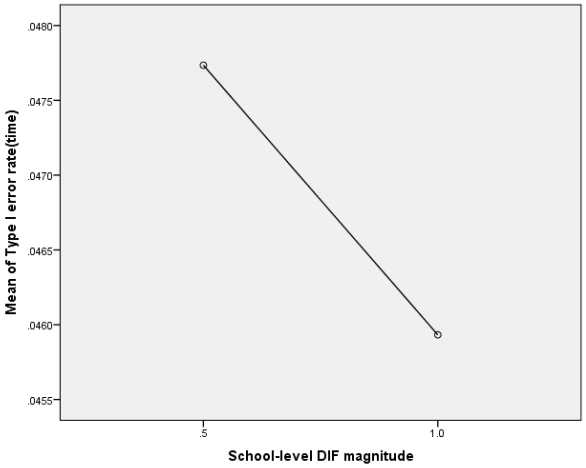

Figure 7. Impact of school-level DIF on time-level Type I error.

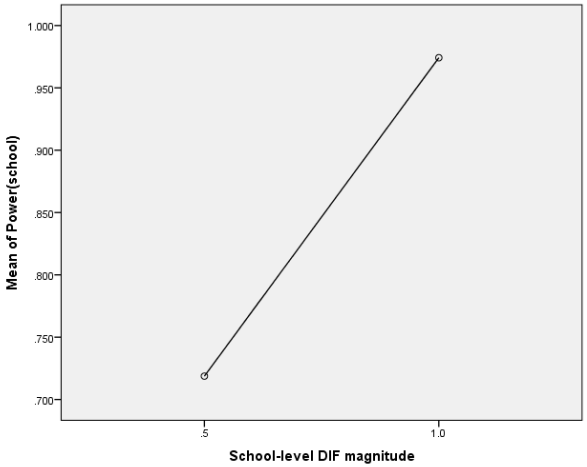

Figure 8. Impact of school-level DIF on school-level power.

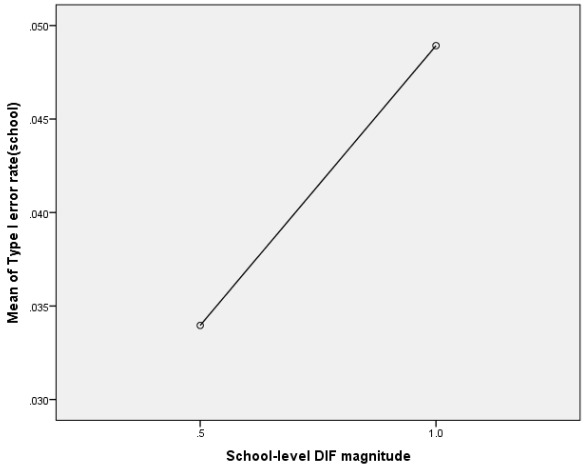

Figure 9. Impact of school-level DIF on school-level Type I error.

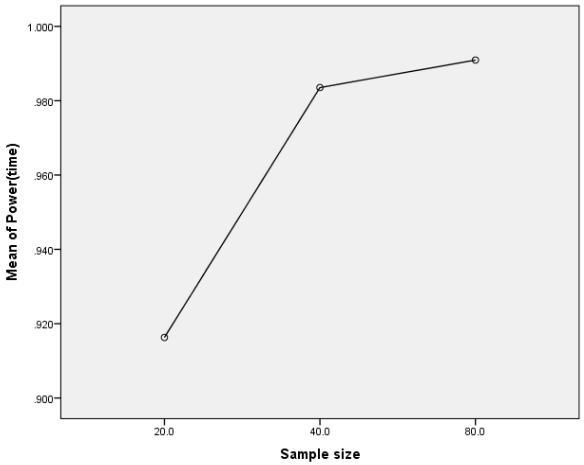

Figure 10. Impact of sample size on time-level power.

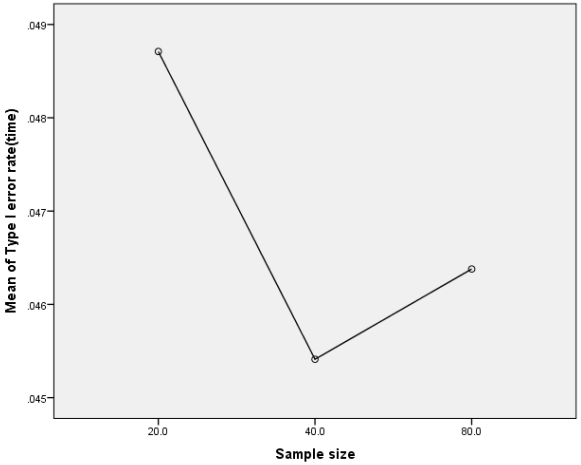

Figure 11. Impact of sample size on time-level Type I error.

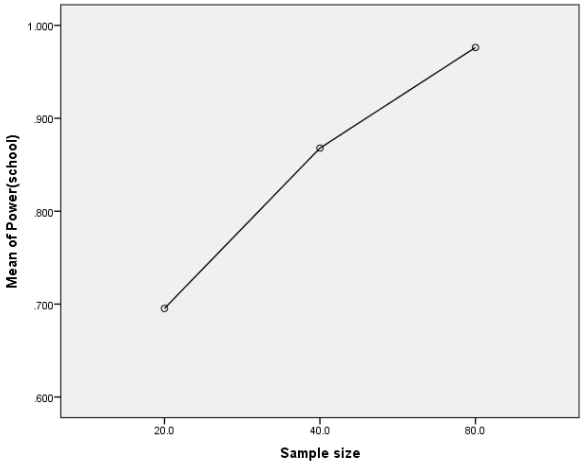

Figure 12. Impact of sample size on school-level power.

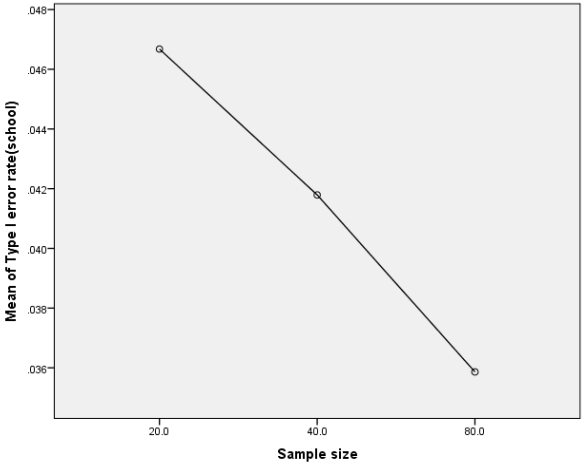

Figure 13. Impact of sample size on school-level Type I error.

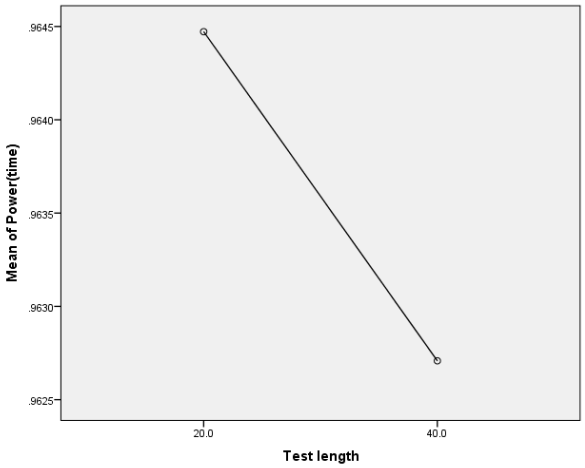

Figure 14. Impact of test length on time-level power.

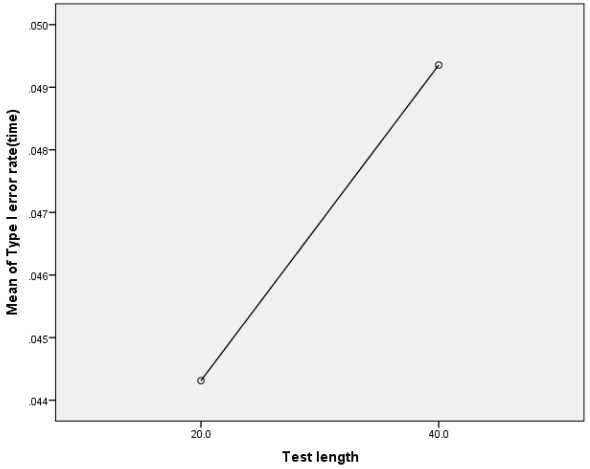

Figure 15. Impact of test length on time-level Type I error.

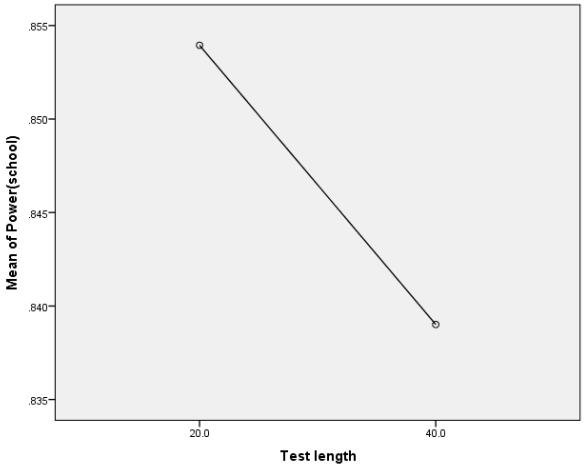

Figure 16. Impact of test length on school-level power.

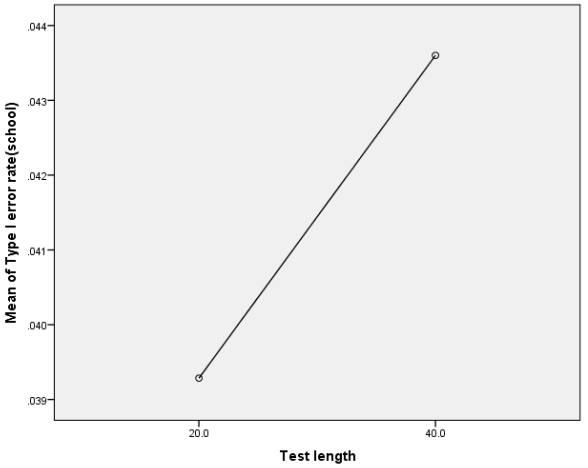

Figure 17. Impact of test length on school-level Type I error.

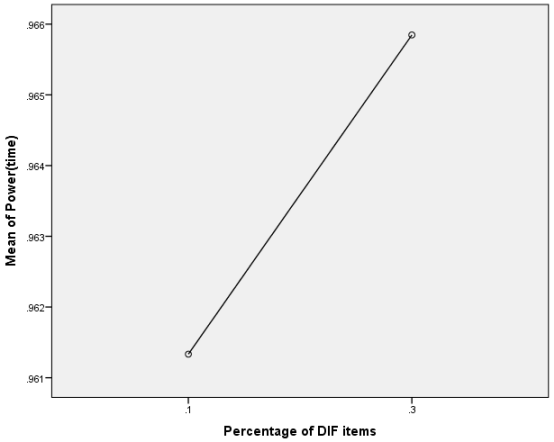

Figure 18. Impact of percentage of DIF items on time-level power.

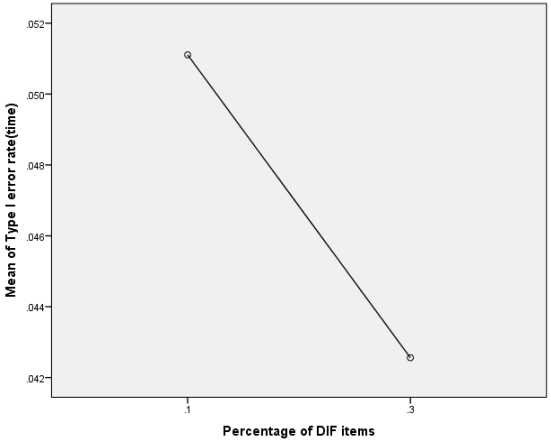

Figure 19. Impact of percentage of DIF items on time-level Type I error.

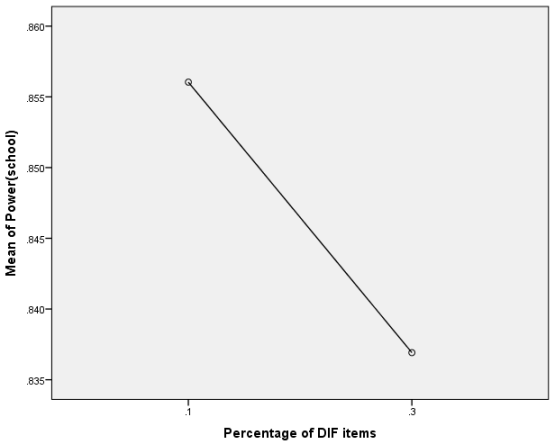

Figure 20. Impact of percentage of DIF items on school-level power.

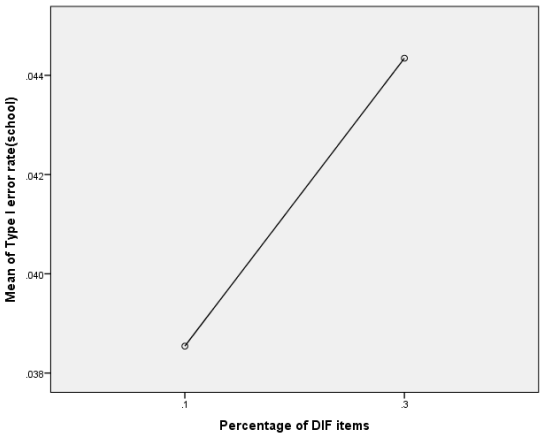

Figure 21. Impact of percentage of DIF items on school-level Type I error.

Figure 10 depicted that time-level power increased as the sample size increased. When the number of schools= 20, 40, and 80, time-level power= 0.916, 0.984, and 0.991, respectively. School-level power also increased as the sample size increased according to Figure 12. When number of schools= 20, 40, and 80, the school-level power= .695, .868, and .976, respectively.

No significant effect was found in the factor of test length. Therefore, the results were not interpreted from Figure 14 to Figure 17.

Figure 18 indicated an increase of Type I error rate at time-level when the percentage of DIF items decreased. When percentage of DIF items= 0.5, the Type I error rate= .051, whereas when percentage of DIF items=1, the Type I error rate= .043.

In short, with a large magnitude of DIF and large sample size, the power of DIF detection at each level using the proposed longitudinal multilevel logistic regression model was high. Practitioners may use this model to detect either time-level or school-level or both-level DIF when it is hypothesized that the magnitudes of DIF and the sample size are large. The average power at time level (0.964) was higher than the average power at school level (0.846), which implied a more powerful DIF detection at the time level.

Only small effects (marginal significant results) of the factors were found on Type I error rates. When the magnitude of DIF at school level was low, a low Type I error rate was found at school level, whereas a low time-level Type I error rate was found when the percentage of DIF items at time-level was high.

Discussion

The present study proposed and investigated a longitudinal multilevel logistic regression model for DIF detection under a variety of simulation conditions. DIF in multilevel can be complicated due to the different types of random effects (Wen, 2014). Two types of random effects were considered in the current study: (a) time-level random effects with a correlation among the errors of a pretest and a posttest, (b) school-level random effects that normally distributed. DIF in multilevel data was explored, assuming that (a) only item difficulty parameters vary across items whereas item discrimination parameter was constrained to 1, (b) the multilevel data meet the invariant item assumption in IRT, (c) the school-level group membership across clusters were fixed.

After the investigation of the Type I error rates and power of detecting DIF in multilevel data, the longitudinal multilevel logistic regression model for DIF detection has been proved to be powerful and accurate in identifying DIF at both the time and school levels. The currently available DIF detection methods, for instance, ANOVA method, the Mantel-Haenszel procedure (Narayanan & Swaminathan, 1994), and the standardized p-difference index, can only detect the existence of DIF, but cannot explain the causes of DIF. The source of DIF is typically assumed to be student characteristics, e.g., student gender and ethnicity. Therefore, the traditional DIF detection is only conducted at the student-level. From a perspective of multilevel modeling, the traditional DIF detection methods assume the impact of DIF is the same across all clusters (Wen, 2014). However, by definition, DIF can occur not only at the student-level, but also levels, such as the time-, class-, and school- level. In the present simulation study, DIF

occurred at both time-level and school-level. At the pretest (the beginning-of-year test), students may not perform as well as at posttest (the mid-of-year test) due to the reasons, such as being nervous for taking the test for the first time, bad weather at that certain time point (e.g. storm or blizzard), or lack of sleep due to the early morning exam. At the posttest, however, the students may not be as nervous for taking the test the second time, or the weather was examinee-friendly, or the students had a good rest or simply felt ready for the test at the posttest time point. The students at the posttest may show better performance even though they have the same ability level as the students at the pretest. At the school-level, students were categorized into high socioeconomic status (SES) group and low SES group. Their performance on the test might be different because of their SES but not their ability levels, which were assumed to be the same in the present simulation study. In a multilevel situation as such, a researcher may be interested in investigating DIF at the time-level and/or the school-level in order to understand the differential performance among students and to modify the items with DIF to fairly evaluate students' abilities. The impacts of DIF at the time-level should not influence DIF detection at the school level, as long as school-level characteristics do not vary within clusters (Ryu, 2014).

In the present study, DIF at both time- level and school- level were shown to be detected successfully with the proposed model. As expected, the magnitudes of time-level and school-level DIF had effects on the DIF detection of these two levels, respectively. The sample size had significant effect on DIF detection at both levels.

More specifically, with large magnitudes of DIF at each level and/or large sample size, the proposed model showed high power at DIF detection (Figure 1 to Figure 12).

The proposed model was the most generalized framework to address DIF from the most sources. The present simulation study has considered the two most common types of random effects in multilevel modeling: (a) the errors are correlated in certain pattern; (b) the errors are independent from each other. More levels can be added in the proposed model under the same framework as long as the type of random effects is appropriately defined and the estimation method is correctly selected. Additionally, the proposed model avoided separate analyses to detect DIF at different levels. Instead, only one model with group membership covariates at each level was utilized, which can be widely applied to situations of (a) detecting time-level DIF only, (b) detecting school-level DIF only, and (c) detecting DIF at both levels at the same time.

The results of the present study were comparable with the previous relevant studies. The magnitude of DIF at each level and the sample size played the most important role in the power of DIF detection (Finch, 2005; Walker et al., 2012; Zumbo, 1999). Relatively large effect was found in the factor of percentage of DIF items (Walker el al, 2012). Linacre (2013) showed, when DIF= 0.5, the smallest sample size for each manifest group must be 300 in order to detect DIF with appropriate power and Type I error rate control. When DIF=1.0, the sample size requirement largely drops to 100 persons per group. The present study confirmed the findings that the larger the magnitude of DIF, the higher the power in DIF detection (Figure 1-Figure 8). The sample sizes of each manifest group in the present study were 250, 500, and 1000, and

results of both power and Type I error rates were in acceptable ranges, with a mean of 0.905 and 0.044, respectively. This was close to the results of Atar (2006), with a mean Type I error rate of 0.0608 and mean power of .8582 across the similar sample size conditions 300, 600, and 1200 in each manifest group.

Practical Implications

The proposed longitudinal multilevel logistic regression model for DIF detection had its advantages that favor the empirical researchers in practice: (a) The model was flexible, easy and efficient to apply in SAS. SAS GLIMMIX statement yields the estimates of all random effects and fixed effects as well as the estimates of variance covariance matrix. Multiple levels could be easily specified in SAS just by including membership covariates at each level. Therefore, there was no need to conduct separate analyses that increase the estimation time and potential Type I error rate. (b) The proposed model can easily include more levels by incorporating more random group effects so that the DIF detection at multiple levels can be conducted simultaneously. The empirical researchers would be able to investigate multiple causes of DIF, better evaluate the students' abilities and modify the items with DIF for future tests. (c) One criticism of the multilevel logistic regression model is that it requires large sample sizes (Hox, 2010; Raudenbush & Anthony, 2001). One possible alternative is to estimate the multilevel logistic regression model with a Markov Chain Monte Carlo (MCMC) simulation (Chaimongkol, 2005) because it works well with small sample sizes (Christensen, Johnson, Branscum, & Hanson, 2010). However, the method would take too long to converge and tend to be time consuming.

Limitations

One major limitation of the present study was the item invariance assumption, which means the items effects are fixed across clusters. In practice, if there is cluster bias of item effects, such bias should be addressed before any DIF analyses. Wen (2014) pointed out if cluster bias existed, one can still conduct DIF analyses using multilevel logistic regression model, although a fixed group membership and random item effects needed to be included to the model. Any further DIF analyses should only be conducted after the cluster bias was tested and addressed.

Another limitation of the present study is, for mathematical simplicity, only 1-PL Rasch IRT model was discussed while item discrimination parameter was constrained to one. Given the complexity of the structure of the proposed longitudinal multilevel model, the estimation of two item parameters will be computationally challenging. Although, Kim et al. (in press) have conducted a 2PL DIF detection at student level in multilevel data with a multilevel MIMIC model. However, because the use of MIMIC model for DIF detection yields high Type I error rates (Finch, 2005; Kim & Yoon, 2011), more empirical studies are needed to determine the feasibility and reliability of the multilevel MIMIC model and multilevel logistic regression model as well as which model is more appropriate to use under what circumstance.

Conclusions

Multilevel DIF have started to be considered recently to detect DIF from different sources. The present study indicated that the proposed model performed well in detecting DIF at time-level with two time points with a correlation of .81 and school-

level with two manifest groups (i.e., the high SES and low SES groups). Compared with the previous relevant studies, the present study managed to detect and explain time-level DIF for the first time, where the correlation of errors of the two time points was considered. The proposed model was easy and flexible to apply in SAS and the estimation consumed reasonably amount of time. The results indicated an overall powerful DIF detection. Type I error rates at each level roughly fell into liberal range of Bradley (1978), 0.025 to 0.075. Consistent with previous study, the magnitude of DIF at each level and the sample was found to be the most important factors for a powerful DIF detection. In general, the time-level detection had higher power than the school-level. Type I error rates were not largely influenced by the factors, although high percentage of DIF items led to low Type I error rate, which was consistent with the results from the previous studies.

SUMMARY AND CONCLUSIONS

There has been growing interest in assessing the students' abilities and holding schools, communities or states accountable for students' performance through high-stakes testing in current educational system. Unfortunately, the traditional single-level assessment methods can only estimate the students' abilities but not the impacts of higher-level clusters, e.g., schools or states. Multilevel analysis makes it possible to treat data in a "multilevel" fashion, taking the variances of higher-level clusters into consideration and extending the measurement from item characteristics and individual abilities to group-level measurements (Bryk & Raudenbush, 2002; Fox, 2005; Goldstein, 1987; Longford, 1993). Multilevel modeling can be combined with Item Response Theory (IRT) to estimate the effects of multilevel covariates on a latent trait. This amalgamation of the two models allows us to investigate and analyze the covariates that affect person ability instead of simply estimating the latent traits (Maier, 2001). Applying the multilevel modeling to logistic regression DIF analyses could examine the causes of DIF by estimating the variances of the random variables.

Study I compared the estimation accuracy of IRT difficulty parameters and abilities estimates of two multilevel IRT modeling (i.e, MR-IRT and MLIRT models). The findings indicated that compared to MR-IRT model, MLIRT model was more accurate in estimating the school level accountability. However, MR-IRT model provided more accurate estimates for individual student abilities, and was more appropriate to use when sample size was small. For both models, the longer the test

length, the more accurate of the estimates. *ICC* did not have the significant impact on the estimation of person abilities or IRT difficulty parameter. But *ICC* played an important role in estimating the school variances of abilities.

For future research, more factors should be investigated to determine which model is more appropriate to use under which circumstance. MLIRT was able to generalize estimates of abilities from the sample to the population by treating the abilities as the random effects. However, the tradeoff of generalizability was less accurate estimates of student abilities. In particular, MLIRT model seemed to have convergence issue of ability estimates when the sample size was small. The impact of sample size on the estimation of abilities as well as item parameters should be further examined. Testing the lower limits of sample size in order to use MLIRT would facilitate practitioners' decision making, in terms of data collection and model selection.

Study II was an application of MLIRT to detect and explain DIF when measuring a certain trait across schools and time points. A longitudinal multilevel DIF detection procedure was proposed by extending the Kamata's MLIRT model to a time level and by including group membership covariates at each level for DIF analyses. In specific, MLIRT model was expressed in the multilevel logistic regression model (Adams, Wilson, & Wu, 1997; Kamata, 2001), and treated the coefficients that are associated with DIF as random effects. As a result, MLIRT was able to estimate the DIF at each level through the estimation of variances of the random effects (Swanminathan & Rogers, 1990). Multiple causes of DIF could be simultaneously analyzed in one multilevel model.

Study II showed that the proposed model performed well in detecting DIF at time-level with a correlation of .81 between the two time points and school-level with two manifest groups (i.e., the high SES and low SES groups). The proposed model filled a major void in education testing, because on top of simply detecting DIF (as what the other DIF detection procedures do), the proposed procedure can also explain the causes of DIF. This type of information will help in conceptualizing the causes of DIF for items and assist the item writers in revising items with DIF.

The proposed model was easy and flexible to apply in SAS and the estimation consumed reasonably amount of time. The results indicated an overall powerful DIF detection, with the Type I error rates at each level roughly falling into liberal range of Bradley (1978), 0.025 to 0.075. The findings were consistent with previous study. The magnitude of DIF at each level and the sample was found to be the most important factors for a powerful DIF detection. In general, the time-level detection had higher power than the school-level. Type I error rates were not largely influenced by the factors, although high percentage of DIF items was found to result in low Type I error rate.

REFERENCES

Adams, R. J., Wilson, M. & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22,* 47-76.

Angoff, W.H. (1982). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-24). Hillsdale, NJ: Erlbaum.

Atar, B. (2006). Differential item functioning analyses for mixed response data using IRT likelihood-ration test, logistic regression, and GLLAMM procedures. *Electronic Theses, Treatises and Dissertations.* Paper 248.

Balluerka, B., Gorostiaga, A., Gómez-Benito, J., & Hidalgo, D. (2010). Use of multilevel logistic regression to identify the causes of differential item functioning. *Psicothema, 22*, 1018-1025.

Bielinski, J., Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple-choice mathematics items administered to national probability samples. *Journal of Educational Measurement, 38*, 51-77.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematics and Statistical Psychology, 31*, 144-152.

Brune, K. D. (2011). *An evaluation of item difficulty and person ability estimation using the multilevel measurement model with short tests and small sample sizes* (Unpublished doctoral dissertation). University of Texas, Austin.

Bryk, A.S., & Raudenbush, S.W. (2002). *Hierarchical linear models for social and behavioural research: Applications and data analysis methods*. Newbury Park, CA: Sage.

Burstein, L. (1980). The analysis of multilevel data in education and evaluation. *Review of Research in Education, 8*, 158-233.

Camilli, G.L., & Shepard, L.A. (1994). Methods for identifying biased test items. Thousand Oakes, CA: Sage.

Carsey, T. & Harden, J. (2014). *Monte Carlo simulation and resampling methods for social science.* Thousand Oaks, CA: Sage.

Chaimongkol, S. (2005). *Modeling differential item functioning (DIF) using multilevel logistic regression models: A Bayesian perspective* (Unpublished doctoral dissertation). Florida State University, Tallahassee.

Cho, S., (2007). *A multilevel mixture model for DIF analysis* (Unpublished doctoral dissertation). University of Georgia, Athens.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items: An NCME instructional module. *Educational Measurement: Issues and Practice, 17*(1), 31-44.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed). Mahwah, NJ: Erlbaum.

Cohen, A. S., Gregg, N., & Deng, M. (2004). *A mixture model analysis of the item level impact of testing accommodations*. Paper presented at the annual meeting of the International Testing Conference, Williamsburg, VA.

Cole, N. S. (1993). History and development of DIF. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 25-29). Hillsdale, NJ: Erlbaum.

Cole, R., Haimson, J., Perez-Johnson, I., & May, H. (2011). *Variability in pretest-posttest correlation coefficients by student achievement level (NCEE reference report 2011-4033).* Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design and analysis,* Occasional paper. Stanford, CA: Stanford Evaluation Consortium, Stanford University.

Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data*. ACT research report series 97-4, Iowa City, IA: ACT.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized and nonlinear approach.* New York, NY: Springer.

Dorans, N.J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszeland standardization. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Edward, C. W. (1993). *Revising SAT verbal items to eliminate Differential Item Functioning*. College Board Report, 93-2, New York, NY: The College Board.

Fan, X., Felsövályi, A., Sivo, S. A., & Keenan, S. C. (2002). *SAS for Monte Carlo studies: A guide for quantitative researchers*, Cary, NC: SAS Institute Inc.

Ferne, T., & Rupp, A.A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges and recommendations. *Language Assessment Quarterly, 4*, 113-148.

Finch, W. H. (2005). The MIMIC model as a method for detecting DIF: comparison with Mantel-Haenszel, SIBTEST and IRT likelihood ratio. *Applied Psychological Measurement*, 29(4), 278-295.

Finch, W. H., & French, B. F. (2011). Estimation of MIMIC model parameters with multilevel data. *Structural Equation Modeling*, 18, 229-252.

Fox, J.P. (2004). Applications of Multilevel IRT Modeling. *School Effectiveness and School Improvement, 15*, 261–280.

Fox, J.P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology, 58*, 145-172.

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multi-level IRT model using Gibbs sampling. *Psychometrika, 66*, 271-288.

Gelman, A., & Hill, J. (2007*). Data analysis using regression and multilevel/hierarchical models.* New York, NY: Cambridge University Press.

Goldstein, H. (1987). Multilevel covariance component models. *Biometrika, 74*, 430-31.

Goldstein, H. (1987). The choice of constraints in correspondence analysis. *Psychometrika, 52*, 207-15.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test item: Comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*, 313-334.

Hedeker, D., Berbaum, M., & Mermelstein, R. (2006). Location-scale models for multilevel ordinal data: Between- and within-subjects variance modeling. *Journal of Probability and Statistical Science, 4*, 1-20.

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mentel-Haenszel procedure. In H. Holland & H.I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Hox, J.J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed*)*. New York, NY: Routledge.

Julian, L. J. (2011). Measures of anxiety. *Arthritis Care & Research.* 63(S11): 467-472. doi:10.1002/acr.20561

Kamata, A. (1998). *Some generalizations of the Rasch model: An application of the hierarchical generalized linear model* (Unpublished doctoral dissertation). Michigan State University, East Lansing.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79-93.

Kamata, A., & Binici, S. (2003, June). *Random-effect DIF analysis via hierarchical generalized linear models.* Paper presented at the International Meeting of the Psychometric Society: The 68st annual meeting of the Psychometric Society, Sardinia, Italy.

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika, 54,* 681-697.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: a comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling, 18,* 212-228.

Kim, E. S., Yoon, M., Wen, Y., Luo, W., & Kwok, O. (in press). Within-level group factorial invariance in multilevel data: Multilevel factor mixture and multilevel MIMIC models, *Structure Equation Modeling*.

Kim, J. (2010). Controlling Type I error rate in evaluating differential item functioning for four DIF methods: Use of three procedures for adjustment of multiple item testing. *Educational Policy Studies Dissertations.* Paper 67.

Kim, S.-H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*, 93-116.

Kreft, I., & de Leeuw, J. (1998). *Introduction to multilevel modeling*. London: Sage.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*, 935-953.

Kulick, E., & Hu, P. G. (1989). *Examining the relationship between differential item functioning and item difficulty.* College Board Report, 89-5; ETS, RR 89-18, New York, NY: College Entrance Examination Board.

Le, V. (1999). *Identifying DIF on the NELS 88: History achievement test*. Report for CRESST, CSE- TR-511, Washington, DC: Office of Education Research and Improvement.

Linacre, J. M. (2013). Differential item functioning DIF sample size nomogram, *Rasch Measurement Transactions*, 26(4), 1391.

Longford, N. T. (1989). Fisher scoring algorithm for variance component analysis of data with multilevel structure. In R. D. Bock (Ed). *Multilevel analysis of educational data* (pp. 297-310). Orlando, FL: Academic Press.

Longford, N. T. (1993). *Random coefficient models*. New York, NY: Oxford University Press.

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics, 26,* 307–330.

Mellenbergh, G.J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*, 300-307.

Millsap, R.E., & Everson, H.T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, *18*, 315-328.

Narayanan, P., & Swaninathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*, 257-274.

Natesan, P. (2007). *Estimation of two-parameter multilevel item response models with predictor variables: simulation and substantiation for an urban school district* (Unpublished doctoral dissertation). Texas A&M University, College Station.

Naumann, A., Hochweber, J., & Hartig, J. (2013). *Modeling instructional sensitivity using a longitudinal multilevel differential item functioning approach.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Padilla, J.L., Pérez, C., & González, A. (1998). La explicación del sesgo en los ítems. *Psicothema, 2*, 481-490.

Pan, T., (2008). *Using the multivariate multilevel logistic regression model to detect DIF: A comparison with HGLM and Logistic Regression DIF detection methods* (Unpublished doctoral dissertation). Michigan State University, East Lansing.

Pastor, D. A. (2003). The use of multilevel response theory modeling in applied research: An illustration. *Applied Measurement in Education, 16*, 223–243.

Pike, G. R. (1990). *The performance of females and males on the ACT-COMP exam: An analysis of Differential Item Functioning using Samejima's Graded model*. Knoxville, TN: Center for Assessment Research and Development.

Pine, S. M. (1977). Applications of item characteristic curve theory to the problem of test bias. In D. J. Weiss (Ed.), *Applications of computerized adaptive testing: Proceedings of a symposium presented at the 18th annual convention of the Military Testing Association* (Research Rep. No. 77-1, pp. 37-43). Minneapolis,

MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Plake, B.S. (1981). An ANOVA methodology to identify biased test items that takes instructional level into account. *Educational and Psychological Measurement, 41,* 365-368.

Potenza, M.T., & Dorans, N.J. (1995). DIF assessment for polytomos scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 1,* 23-37.

Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495-502.

Raju, N. S., Bode, R. K., & Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel statistic to detect differential item functioning. *Applied Measurement in Education, 2,* 1-13.

Raju, N.S., Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19,* 353-368.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35,* 543-568.

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes*, *41*, 221-250.

Rijmen, F., Tuerlinckx, F., De Boeck, P. & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8,* 185–205.

Rock, D. A., Pollack, J. M., & Quinn, P. (1995). *Psychometric report for NELS: 88 base year through second follow-up* (NCES report 95-382). Washington, DC: U.S. Department of Education.

Roussos, L.A., & Stout, W.F. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355- 371.

Ryu, E. (2014). Factorial invariance in multilevel confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology*, 67(1), 172-194.

SAS Institute Inc. (2011). *SAS/STAT® 9.3 User's Guide.* Cary, NC: SAS Institute Inc.

Schabenberger, O., & Pierce, F. J. (2002). *Contemporary Statistical Models for the Plant and Soil Sciences,* Boca Raton, FL: CRC Press.

Shealy, R., & Stout, W.F. (1993a). An item response theory for test bias. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197-239). Hillsdale, NJ: Erlbaum.

Shealy, R., & Stout, W. F. (1993b). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*, 159-194.

Stine-Morrow, E. A. L., Millinder, L., Pullara, O., & Herman, E. (2001). Patterns of resource allocation are reliable among younger and older readers. *Psychology & Aging, 16*, 69–84.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361–370.

Swanson, D. B., Clauser, B.E., Case, S.M., Nungester, R.J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics, 27*, 53-75.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. Psychometrika, *47*, 175-186.

Thissen, D., Steinberg, L., &Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Van der Leeden, R. (1998). Multilevel analysis of repeated measures data. *Quality and Quantity, 32*, 15-29.

Van den Noortgate, W., De Boeck, P. (2005). Assessing and Explaining Differential Item Functioning Using Logistic Mixed Models. *Journal of Educational and Behavioral Statistics, 30,* 443.

Walker, C. M., Zhang, B., Banks, K., & Cappaert, K. (2012). Establishing effect size guidelines for interpreting the results for differential bundle functioning analyses using SIBTEST. *Educational and Psychological Measurement,* 72(3), 415-434.

Wang, W., & Shih, C. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement,* 34(3), 166-180.

Wen, Y. (2014). DIF analyses in multilevel data: identification and effects on ability estimates. *Theses and Dissertations*. Paper 573.

Whitmore, M. L., & Schumacker, R. E. (1999). A comparison of logistic regression and analysis of variance differential item functioning detection methods. *Applied Psychological Measurement, 59,* 910-927.

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement,* 33, 42-57.

Wright, B. D. (1977). Solving measurement problems with the Rash model. *Journal of Educational Measurement, 14,* 97-116.

Zhu, X. S., Rupp, A. A. & Gao, J. (2011). Differential item functioning analyses in large-scale educational surveys: Key concepts and modeling approaches for secondary analysts. *Journal of Research in Education Sciences, 56,* 91-127.

Zumbo, B. D. (1999). *A handbook on the theory and methods for differentia litem functioning: Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B.D., & Gelin, M.N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological / community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies, 5,* 1-23.

Zwaan, R. A. (1994). Effect of genre expectations on text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 920–933.

Zwick, R. & Ercikan, K. (1989). Analaysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement, 26,* 55-66.